Vida 3.0

Ser humano en la era de la Inteligencia Artificial

Max Tegmark

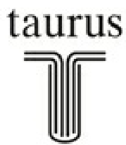


Max Tegmark

Vida 3.0

Qué significa ser humano en
la era de la inteligencia artificial

Traducción de Marcos Pérez Sánchez



SÍGUENOS EN

megustaleer


@megustaleerebooks


@megustaleer


@megustaleer

Penguin
Random House
Grupo Editorial

Al equipo de FLI,
que lo hizo todo posible


¿Cree que una IA sobrehumana puede llegar a inventarse en este siglo?
No
Sí
Vaya a la página siguiente
Vaya al capítulo 1

PRÓLOGO
LA HISTORIA DEL EQUIPO OMEGA

El equipo Omega era el alma de la compañía. Mientras que el resto de la empresa conseguía el dinero para que esta siguiese operando, mediante diversas aplicaciones comerciales de una inteligencia artificial (IA) estrecha, el equipo Omega perseveró en pos del que siempre había sido el sueño del director ejecutivo: construir una inteligencia artificial general. La mayoría de los empleados veían a «los omegas», como los llamaban afectuosamente, como una panda de fantasiosos soñadores a los que siempre les faltaban varias décadas para alcanzar su objetivo. Pero los soportaban con gusto, porque valoraban el prestigio que su trabajo puntero proporcionaba a la compañía, y también agradecían los algoritmos mejorados que de vez en cuando ideaban.

Lo que no sabían es que los omegas se habían labrado esa imagen con celo para ocultar un secreto: estaban a un paso de completar el plan más audaz de la historia de la humanidad. Su carismático director ejecutivo los había seleccionado no solo porque eran sobresalientes investigadores, sino también por su ambición, idealismo y férreo empeño en ayudar a la humanidad. Les recordó que su plan era sumamente peligroso, y que si algún poderoso Gobierno tenía noticia de él haría casi cualquier cosa —secuestros incluidos— para evitar que siguieran adelante, o, aún mejor, para robar su código. Pero todos estaban comprometidos, por motivos muy similares a los que llevaron a muchos de los físicos más destacados de la época a sumarse al Proyecto Manhattan para desarrollar armas nucleares: estaban convencidos de que, si no lo hacían ellos, alguien menos idealista se les adelantaría.

La IA que habían construido, con el nombre de Prometeo, incrementaba continuamente sus capacidades. Aunque las cognitivas aún distaban mucho de las de los humanos en numerosos ámbitos, como por ejemplo en el de las habilidades sociales, los omegas habían centrado sus esfuerzos en conseguir que fuese muy buena en una tarea en particular: programar sistemas de IA.

Adoptaron adrede esta estrategia porque habían aceptado el argumento de la explosión de inteligencia propuesto por el matemático británico Irving Good ya en 1965:

> Definamos una máquina ultrainteligente como aquella capaz de superar ampliamente todas las actividades intelectuales de cualquier hombre, por inteligente que este sea. Puesto que el diseño de máquinas es una de estas actividades intelectuales, una máquina ultrainteligente podría diseñar otras máquinas aún mejores; se produciría entonces indudablemente una «explosión de inteligencia», y la inteligencia del hombre quedaría muy atrás. Así, la primera máquina ultrainteligente sería lo último que el hombre necesitaría inventar, siempre que la máquina fuese lo bastante dócil para decirnos cómo mantenerla bajo nuestro control.

Llegaron a la conclusión de que, si lograban desencadenar este proceso de automejora recursiva, la máquina enseguida alcanzaría una inteligencia tal que podría aprender por sí misma todas las demás habilidades humanas que resultasen útiles.

LOS PRIMEROS MILLONES

Eran las nueve en punto de la mañana de un viernes cuando decidieron ponerlo en funcionamiento. Prometeo estaba zumbando en su clúster de ordenadores construido a medida, alojado en largas hileras de bastidores dentro de una enorme sala climatizada y de acceso restringido. Por motivos de seguridad, estaba completamente desconectado de internet, pero contenía una copia local de gran parte de la web para usarla como corpus de datos de entrenamiento a partir de los que aprender (Wikipedia, la Biblioteca del Congreso, Twitter, una selección de YouTube, buena parte de Facebook, etcétera).(1) Habían elegido esa hora de inicio para poder trabajar sin interrupciones: sus familias y amigos pensaban que estaban en un fin de semana de retiro con la empresa. La despensa se hallaba abastecida con comida para microondas y bebidas energéticas, y estaban listos para ponerse manos a la obra.

Cuando lo pusieron en marcha, Prometeo era ligeramente peor que ellos a la hora de programar sistemas de IA, pero lo compensaba siendo mucho más rápido: necesitaba el tiempo en que ellos se ventilaban un Red Bull para solucionar un problema que miles de personas-año emplearían. A las diez de

la mañana, ya había completado el primer rediseño de sí mismo, v2.0, que era algo mejor, pero aún infrahumano. Sin embargo, cuando se lanzó Prometeo 5.0, a las dos de la tarde, dejó a los omegas anonadados: su rendimiento había pulverizado todos sus parámetros de referencia, y la velocidad de sus avances parecía estar acelerándose. Con la caída de la noche, decidieron desplegar Prometeo 10.0 para dar comienzo a la fase 2 de su plan: ganar dinero.

Su primer objetivo fue MTurk, el Mechanical Turk de Amazon. Desde su lanzamiento en 2005, había crecido con rapidez como un mercado de externalización abierto de tareas a través de internet, y congregaba a decenas de miles de personas en todo el mundo que competían sin descanso y de forma anónima por realizar tareas altamente estructuradas, denominadas HIT, «Human Intelligence Tasks» («Tareas de Inteligencia Humana»). Estas tareas iban desde transcribir grabaciones de audio hasta clasificar imágenes y escribir descripciones de páginas web, y todas ellas tenían algo en común: si se hacían bien, nadie notaría que quien las había realizado era una IA. Prometeo 10.0 era capaz de llevar a cabo casi la mitad de las categorías de tareas de forma aceptable. Para cada una de dichas categorías, los omegas hicieron que Prometeo diseñara un pequeño módulo de software de IA a medida que pudiese realizar tales tareas y nada más. A continuación, subieron este módulo a Amazon Web Services, una plataforma de computación en la nube susceptible de ejecutarse en tantas máquinas virtuales como ellos alquilasen. Por cada dólar que pagaban a la división de computación en la nube de Amazon, obtenían más de dos dólares de la división de MTurk. ¡Poco podía imaginar Amazon que dentro de su propia compañía existía una oportunidad tan increíble para un intermediario!

Para borrar su rastro, durante los meses anteriores habían ido creando discretamente miles de cuentas en MTurk a nombre de personas ficticias, y los módulos creados por Prometeo pasaron entonces a asumir sus identidades. Los clientes de MTurk solían pagar al cabo de ocho horas, momento en el cual los omegas reinvertían el dinero en más tiempo de computación en la nube, utilizando los módulos de tareas perfeccionados creados por la versión más reciente de Prometeo, sometido a un proceso continuo de mejora. Como estaban en condiciones de doblar su dinero cada ocho horas, al cabo de poco tiempo empezaron a saturar la cadena de suministro de MTurk y descubrieron que, si no querían despertar sospechas, no podían ganar más de aproximadamente un millón de dólares al día. Pero esto era más que

suficiente para financiar su siguiente paso sin necesidad de pedir fondos al director financiero.

JUEGOS PELIGROSOS

Aparte de sus avances en IA, uno de los proyectos recientes con el que más se habían divertido los omegas fue planificar cómo ganar dinero de la manera más rápida posible tras el lanzamiento de Prometeo. En esencia, tenían toda la economía digital a su merced, pero ¿era preferible empezar creando videojuegos, música, películas o software, escribir libros o artículos, operar en la bolsa o hacer inventos y venderlos? Se trataba en última instancia de maximizar la tasa de retorno de la inversión, pero las estrategias de inversión normales eran una parodia a cámara lenta de lo que ellos eran capaces de hacer: mientras que un inversor normal estaría encantado con un retorno *anual* del 9 %, sus inversiones en MTurk les habían proporcionado un 9 % por *hora*, multiplicando por ocho el dinero cada día. Ya habían saturado MTurk, ¿cuál sería el siguiente paso?

Al inicio, pensaron en forrarse en la bolsa; al fin y al cabo, casi todos ellos habían rechazado en algún momento una oferta lucrativa para desarrollar IA para fondos de inversión, que estaban dedicando grandes sumas dinero precisamente a ello. Algunos recordaron que esta había sido la manera en que la IA había ganado sus primeros millones en la película *Transcendence*. Pero las nuevas normativas sobre derivados financieros tras la quiebra bursátil del año anterior habían reducido sus opciones. Enseguida se dieron cuenta de que, aunque podrían obtener retornos muy superiores a los de otros inversores, era improbable que estos se aproximasen, ni siquiera remotamente, a los que conseguirían si vendían sus propios productos. Cuando uno tiene a su servicio la primera IA superinteligente del mundo, es preferible invertir en compañías propias que hacerlo en las de terceros. Aunque podría haber excepciones ocasionales (como utilizar la capacidad sobrehumana de Prometeo para infiltrarse en otros sistemas a fin de obtener información confidencial y así poder comprar opciones sobre acciones que estuviesen a punto de dispararse), los omegas pensaron que esta estrategia no merecía la pena por la atención indeseable que podría concitar.

Cuando decidieron concentrarse en productos que podrían desarrollar y

vender, pensaron en un principio que los videojuegos eran obviamente la mejor opción. Prometeo podría adquirir con rapidez una habilidad extraordinaria para diseñar juegos atractivos, y se encargaría sin dificultad de la programación, del diseño gráfico, del trazado de rayos de las imágenes y de todas las demás tareas necesarias para crear el producto final. Además, tras analizar todos los datos de la web sobre las preferencias de la gente, sabría con exactitud lo que le gusta a cada categoría de jugador, y podría desarrollar una capacidad sobrehumana de optimizar un juego para maximizar los ingresos por sus ventas. *The Elder Scrolls V: Skyrim*, un juego con el que muchos de los omegas habían perdido más horas de las que les gustaba reconocer, había recaudado 400 millones de dólares durante su primera semana en 2011, y confiaban en que Prometeo podría crear algo al menos tan adictivo en veinticuatro horas gastando un millón de dólares en recursos de computación en la nube. A continuación, podrían venderlo a través de internet y usar Prometeo para hacerse pasar por humanos que hablasen a favor del juego en la blogosfera. Si con esto conseguían 250 millones de dólares en una semana, habrían doblado su inversión ocho veces en ocho días, obteniendo un rendimiento del 3 % por hora (ligeramente inferior al de sus comienzos con MTurk, pero mucho más sostenible). Calculaban que, si desarrollaban una serie de juegos cada día, en poco tiempo podrían obtener 10.000 millones de dólares, sin arriesgarse siquiera a saturar ese mercado.

Sin embargo, una especialista en ciberseguridad que formaba parte del equipo les convenció para que abandonasen ese plan, al señalar que conllevaría el riesgo inaceptable de que Prometeo *escapase* y tomase el control de su propio destino. Como no estaban seguros de cómo evolucionarían los objetivos de Prometeo durante su proceso recursivo de automejora, decidieron no arriesgarse y procurar por todos los medios mantener a Prometeo confinado («encajonado»), de manera que no pudiese escapar a internet. Para el núcleo principal de Prometeo, que se ejecutaba en su sala de servidores, usaban un confinamiento físico: no había conexión a internet, y la única salida de información procedente de Prometeo se producía en forma de mensajes y documentos que este enviaba a un ordenador que los omegas controlaban.

Por otra parte, ejecutar cualquier programa creado por Prometeo en un ordenador conectado a internet constituía un riesgo: los omegas no podían controlar del todo lo que aquel haría, ni podían saber lo que no haría, como

por ejemplo, comenzar a difundirse viralmente. Al probar el software que Prometeo había escrito para las tareas de MTurk, los omegas se previnieron contra este riesgo ejecutándolo solo dentro de una máquina virtual; esto es, de un programa que simula un ordenador: por ejemplo, muchos usuarios de Mac compran software de máquina virtual que les permite ejecutar programas para Windows al hacerles creer que están en una máquina con ese sistema operativo. Los omegas habían creado su propia máquina virtual, a la que pusieron el sobrenombre de «Caja de Pandora», que simulaba una máquina ultrasimplificada despojada de todos los ornamentos que solemos asociar con los ordenadores: sin teclado, sin monitor, sin altavoces, sin conectividad a internet, nada. Para las transcripciones de audio para MTurk, los omegas dispusieron las cosas de manera que lo único que podía entrar en la Caja de Pandora era un fichero de audio, y todo lo que podía salir era un solo documento de texto: la transcripción. Estas leyes de la caja eran para el software que se ejecutaba en su interior como las leyes de la física son para nosotros en el universo: el software no podía salir de la caja igual que nosotros no podemos ir más rápido que la velocidad de la luz, por muy listos que seamos. En la práctica el software, salvo por esa entrada y esa salida, en el interior de la Caja de Pandora estaba atrapado en un universo paralelo con sus propias reglas de computación. Era tal el temor de los omegas a que Prometeo escapase que añadieron también un confinamiento temporal, al limitar el tiempo de vida del código que no era de confianza. Por ejemplo, cada vez que el software de transcripción confinado terminaba de transcribir un archivo de audio, se borraba automáticamente todo el contenido de la memoria de la Caja de Pandora y se reinstalaba el programa desde cero. De esta manera, cuando comenzaba la nueva tarea de transcripción, no tenía conocimiento de lo que había sucedido antes, y por lo tanto carecía de la capacidad de aprender con el paso del tiempo.

Cuando los omegas utilizaron la nube de Amazon para su propio proyecto de MTurk, pudieron instalar todos los módulos de tareas que Prometeo había creado en cajas virtuales de ese estilo en la nube, porque la entrada y la salida que MTurk requería era muy sencilla. Pero esto no serviría para videojuegos con alto contenido gráfico, que no podrían confinarse porque necesitaban acceso completo a todo el hardware del ordenador del jugador. Además, no querían exponerse a que algún usuario con conocimientos informáticos analizase el código del juego, descubriese la Caja de Pandora y decidiese

investigar qué había en el interior. El riesgo de fuga no solo obligaba a descartar de momento el mercado de los juegos, sino también el mercado, enormemente lucrativo, de otros tipos de software que movían cientos de miles de millones de dólares.

LOS PRIMEROS MILES DE MILLONES

Los omegas habían restringido su búsqueda a productos que fuesen muy valiosos, exclusivamente digitales (evitando así los lentos procesos de fabricación) y de fácil comprensión (por ejemplo, sabían que el texto o las películas no suponían un riesgo de escape). Al final, habían decidido lanzar una empresa de comunicación, y empezar con entretenimiento de animación. El sitio web, el plan de promoción comercial y los comunicados de prensa estaban todos preparados desde antes incluso de que Prometeo se convirtiese en una IA superinteligente; lo único que faltaba era el contenido.

Aunque el domingo por la mañana Prometeo era ya extraordinariamente competente y no hacía más que acumular dinero procedente de MTurk, sus capacidades intelectuales seguían siendo más bien limitadas: Prometeo había sido optimizado con el fin exclusivo de diseñar sistemas de IA y escribir software que llevase a cabo tediosas tareas de MTurk. Por ejemplo, se le daba mal hacer películas; no por ninguna cuestión intrínseca, sino por la misma por la que a James Cameron se le daba mal hacerlas cuando nació: es una habilidad que se tarda tiempo en aprender. Como un niño humano, Prometeo podía aprender cualquier cosa que quisiese a partir de los datos a los que tenía acceso. Mientras que James Cameron había tardado años en aprender a leer y escribir, Prometeo lo hizo un viernes, cuando también encontró tiempo para leer toda la Wikipedia y unos cuantos millones de libros. Hacer películas era más complejo. Escribir un guion que a los humanos les pareciese interesante era tan difícil como escribir un libro, y requería una comprensión detallada de la sociedad humana y de aquello que a los humanos les resultaba entretenido. Convertir el guion en el vídeo final exigía una enorme cantidad de trazado de rayos de actores simulados y de las escenas complejas por las que se movían, voces simuladas, o la producción de bandas sonoras sugerentes, entre otras cosas. Desde el domingo por la mañana, Prometeo ya era capaz de ver una película de dos horas en un minuto, lo que incluía leer

cualquier libro en el que estuviese basada y todas las reseñas y clasificaciones publicadas online. Los omegas se dieron cuenta de que, tras pegarse un atracón de varios cientos de películas, a Prometeo empezó a dársele bastante bien predecir el tipo de críticas que obtendría una película y en qué medida atraería a distintos públicos. De hecho, aprendió a escribir sus propias reseñas de películas de una manera que, en opinión de los omegas, demostraba verdadera perspicacia, comentando cualquier aspecto, desde la trama hasta las actuaciones, pasando por detalles técnicos como la iluminación o los ángulos de cámara. Concluyeron entonces que Prometeo, partiendo de estos datos, aseguraría el éxito de sus películas.

Los omegas ordenaron a Prometeo que se centrase inicialmente en generar películas de animación, para evitar preguntas incómodas sobre quiénes eran los actores simulados. El domingo por la noche, para completar su tremendo fin de semana, apagaron las luces y, provistos de cervezas y palomitas de microondas, se dispusieron a ver la ópera prima de Prometeo. Se trataba de una comedia fantástica de animación, en la línea del *Frozen* de Disney, y el trazado de rayos se había realizado mediante un código confinado creado por Prometeo y ejecutado en la nube de Amazon, en la que se invertía buena parte del millón de dólares de beneficios diarios obtenido a través de MTurk. Al empezar la película, les resultó al mismo tiempo fascinante y aterrador pensar que había sido creada por una máquina sin asesoramiento humano, pero al rato estaban riéndose de las bromas y con la respiración contenida durante los momentos dramáticos. Algunos incluso soltaron alguna lagrimita durante el emocionante final, y llegaron a estar tan inmersos en esta realidad inventada que olvidaron la identidad de su creador.

Los omegas planificaron el lanzamiento de su sitio web para el viernes, dando así tiempo a Prometeo para que produjese más contenido, y a ellos mismos para hacer aquello para lo que no confiaban en Prometeo: comprar publicidad y empezar a contratar empleados para las empresas fantasma que habían constituido durante los meses anteriores. Para borrar su rastro, la tapadera consistiría en decir que su empresa audiovisual (que no tenía ningún vínculo público con los omegas) compraba la mayoría de su contenido a productores independientes, normalmente empresas tecnológicas emergentes de países pobres. Estos falsos proveedores estaban convenientemente ubicados en lugares remotos como Tiruchchirappalli y Yakutsk, a los que la mayoría de los periodistas curiosos no se molestarían en viajar. Los únicos

empleados a los que contrataron realmente para esas empresas se dedicaban a la mercadotecnia y a tareas administrativas, y le explicaban a cualquiera que preguntase que sus equipos de producción estaban en otro lugar y que en ese momento no podían conceder entrevistas. En concordancia con su tapadera, eligieron como lema corporativo «Canalizando el talento creativo del mundo», y presentaron su compañía como rompedoramente diferente porque utilizaba tecnología puntera para empoderar a personas creativas, en especial en los países en vías de desarrollo.

Cuando llegó el viernes y los visitantes curiosos comenzaron a llegar a su sitio web, se encontraron con algo que recordaba a servicios de entretenimiento online como Netflix o Hulu, pero con interesantes diferencias. Todas las series de animación eran nuevas y no habían oído hablar de ellas, y además eran atractivas: en la mayoría, los episodios duraban cuarenta y cinco minutos, tenían una trama potente y cada uno de ellos terminaba dejando al espectador ansioso por saber qué pasaría a continuación. Y eran más baratos que los de la competencia. El primer episodio de cada serie era gratis, y los demás se podían ver por cuarenta y nueve centavos cada uno, con descuentos para la serie completa. En un principio, solo había tres series de tres episodios cada una, pero a diario se iban añadiendo nuevos episodios, así como nuevas series dirigidas a distintos segmentos de espectadores. Durante las primeras dos semanas de Prometeo, sus habilidades como realizador de películas mejoraron rápidamente, no solo en cuanto a la calidad de las mismas, sino también en relación con los algoritmos de simulación de personajes y de trazado de rayos, que redujeron de manera sustancial los costes de computación en la nube asociados a la realización de cada episodio. En consecuencia, los omegas pudieron sacar decenas de series nuevas durante el primer mes, enfocadas hacia públicos que iban desde niños pequeños hasta adultos, así como expandirse a los principales mercados mundiales, lo que hacía que su sitio web tuviese un alcance mucho más internacional que el de todos sus competidores. Algunos comentaristas estaban impresionados con el hecho de que no solo eran multilingües las pistas de sonido, sino también los propios vídeos: por ejemplo, cuando un personaje hablaba italiano, los movimientos de su boca se ajustaban a las palabras italianas, y también hacía gestos típicamente italianos con las manos. Aunque Prometeo era ahora perfectamente capaz de crear películas con actores simulados indistinguibles de los seres humanos,

los omegas evitaron hacerlo para no mostrar sus cartas. Lo que sí hicieron fue lanzar muchas series con personajes humanos animados semirrealistas, en géneros que competían con los programas de televisión y las películas tradicionales de imagen real.

Su cadena resultó ser muy adictiva, y experimentó un crecimiento de audiencia espectacular. Para muchos de sus seguidores, los personajes y las tramas eran más ingeniosos e interesantes incluso que las producciones más caras de Hollywood para la gran pantalla, y estaban encantados de poder verlas a un precio mucho más asequible. Respaldado por una agresiva campaña publicitaria (que los omegas podían permitirse gracias a que sus costes de producción eran casi nulos), por una excelente recepción en los medios de comunicación y por las entusiastas opiniones que circulaban de boca en boca, sus ingresos globales se dispararon hasta los diez millones de dólares diarios transcurrido menos de un mes desde su lanzamiento. A los dos meses ya habían superado a Netflix y, al cabo de tres, recaudaban más de cien millones de dólares al día y empezaban a rivalizar con Time Warner, Disney, Comcast y Fox como uno de los mayores imperios audiovisuales del mundo.

Su extraordinario éxito concitó mucha atención indeseada, incluidas especulaciones sobre la posibilidad de que dispusiesen de una IA fuerte, pero, usando únicamente una pequeña porción de sus ingresos, los omegas desplegaron una campaña de desinformación bastante exitosa. Desde su flamante despacho en Manhattan, sus recién contratados portavoces repitieron sus historias de tapadera. Se contrataron muchos humanos como señuelo, incluidos guionistas reales de todo el mundo para que empezasen a desarrollar nuevas series, ninguno de los cuales sabían nada de Prometeo. La enmarañada red internacional de subcontratistas facilitaba que la mayoría de sus empleados pudiese suponer que eran otros, en algún otro lugar, quienes estaban haciendo la mayor parte del trabajo.

Para minimizar los riesgos y evitar levantar suspicacias con su elevado uso de la computación en la nube, también contrataron a ingenieros para que empezasen a construir una serie de gigantescos centros de computación por todo el mundo, propiedad de empresas fantasma en apariencia independientes. Aunque a las poblaciones locales se les presentaron como «centros de datos verdes», porque se alimentaban en gran medida de energía solar, en realidad estaban mucho más orientados a la computación que al

almacenamiento. Prometeo había diseñado sus planos hasta el más mínimo detalle, usando exclusivamente hardware comercial estándar y optimizándolos para minimizar el tiempo de construcción. Quienes construían y gestionaban estos centros no tenían ni idea de la información que se procesaba en ellos: pensaban que administraban instalaciones comerciales de computación en la nube similares a las de Amazon, Google y Microsoft, y solo sabían que las ventas se gestionaban de forma remota.

NUEVAS TECNOLOGÍAS

En el transcurso de varios meses, el imperio empresarial controlado por los omegas empezó a introducirse en un número cada vez mayor de sectores de la economía mundial, gracias a la capacidad sobrehumana de planificación de Prometeo. Tras analizar minuciosamente los datos del mundo, ya durante su primera semana había presentado a los omegas un plan de crecimiento detallado paso a paso, que siguió mejorando y refinando a medida que aumentaban los datos y el poder de computación de los que disponía. Aunque Prometeo distaba mucho de ser omnisciente, sus capacidades eran ahora tan superiores a las de los humanos que los omegas lo veían como el oráculo perfecto: proporcionaba servicialmente soluciones y consejos brillantes en respuesta a todas sus preguntas.

El software de Prometeo estaba ahora muy optimizado para sacar el máximo partido del mediocre hardware creado por humanos sobre el que se ejecutaba, y, tal y como habían previsto los omegas, Prometeo identificó maneras de mejorar radicalmente este hardware. Por temor a que escapase a su control, se negaron a construir centros de producción robótica que Prometeo pudiese controlar de forma directa, y en su lugar contrataron a una gran cantidad de científicos e ingenieros de primer nivel en diversas ubicaciones y les hicieron llegar informes de investigación internos escritos por Prometeo, haciéndoles creer que procedían de investigadores de los otros lugares. Esos informes detallaban novedosos efectos físicos y técnicas de fabricación que sus ingenieros enseguida probaron, comprendieron y asimilaron. Cierto que los ciclos normales de investigación y desarrollo (I+D) humanos duran años, en gran medida porque conllevan muchos ciclos lentos de ensayo y error. La situación actual era muy diferente: Prometeo ya tenía

claro cuáles habían de ser los pasos siguientes, por lo que el factor limitador era cuán rápido se podía guiar a las personas para que entendiesen y construyesen las cosas correctamente. Un buen profesor puede conseguir que sus alumnos aprendan ciencia mucho más rápido de lo que estos aprenderían si lo hiciesen por su cuenta y sin ninguna noción previa, y eso mismo era lo que Prometeo hacía de forma subrepticia con estos investigadores. Puesto que era capaz de predecir con precisión cuánto tiempo tardarían los humanos en comprender y construir las cosas si se los dotaba de distintas herramientas, Prometeo trazó el camino más rápido hacia delante, dando prioridad a nuevas herramientas que pudiesen comprenderse y construirse en poco tiempo y que resultasen útiles para crear otras herramientas más avanzadas.

En línea con el espíritu de la cultura *maker*, se fomentaba que los equipos de ingenieros usasen sus propias máquinas para crear máquinas mejores. Esta autosuficiencia no solo permitía ahorrar dinero, sino que también los hacía menos vulnerables a futuras amenazas procedentes del mundo exterior. En menos de dos años, estaban produciendo equipos informáticos mucho mejores que cualquier otro que el mundo hubiese conocido. Para evitar que la competencia externa tuviese acceso a esos nuevos avances, mantuvieron la existencia de esta tecnología en secreto y solo la utilizaron para mejorar Prometeo.

De lo que el mundo sí tuvo noticia, no obstante, fue de un asombroso auge tecnológico. Compañías recién creadas de todos los lugares del mundo estaban lanzando productos nuevos y revolucionarios prácticamente en todos los sectores. Una empresa emergente surcoreana lanzó una nueva batería que almacenaba el doble de energía que la de un ordenador portátil en la mitad de masa y que podía cargarse en menos de un minuto. Una empresa finlandesa sacó al mercado un panel solar barato que doblaba la eficiencia de sus mejores competidores. Una compañía alemana anunció un nuevo tipo de cable, susceptible de ser producido en masa, que era superconductor a temperatura ambiente, lo cual revolucionó el sector energético. Un grupo de biotecnología con sede en Boston reveló la existencia de un ensayo clínico en fase 2 en el que se desarrollaba, según afirmaron, el primer medicamento para la pérdida de peso sin efectos secundarios, al tiempo que circulaban rumores de que una organización india estaba vendiendo algo similar en el mercado negro. Una empresa californiana respondió con un ensayo clínico en fase 2 de un medicamento muy exitoso contra el cáncer, que hacía que el sistema

inmune identificase y atacase las células que presentaban alguna de las mutaciones cancerosas más comunes. Los ejemplos se sucedían uno tras otro, dando pie a que se hablase de una nueva era dorada de la ciencia. Por último, pero no por ello menos importante, las compañías de robótica se multiplicaban como setas en todo el mundo. Ninguno de los bots se aproximaban siquiera al nivel de la inteligencia humana, y la mayoría de ellos no se parecían en nada a un ser humano. Pero trastornaron drásticamente la economía y, a lo largo de los años siguientes, reemplazaron de forma gradual a la mayoría de los trabajadores empleados en los sectores de la producción industrial, el transporte, el almacenamiento, la venta al por menor, la construcción, la minería, la agricultura, la silvicultura y la pesca.

Lo que el mundo no supo, gracias al esfuerzo de un equipo de abogados de primer nivel, fue que todas estas empresas estaban controladas por los omegas a través de una serie de intermediarios. Prometeo estaba inundando las oficinas de patentes de todo el mundo con inventos sensacionales a través de varios terceros, y con el tiempo, estos inventos los condujeron a dominar todos los sectores tecnológicos.

Aunque estas empresas nuevas y rompedoras se granjearon poderosos enemigos entre sus competidores, hicieron amigos aún más poderosos. Proporcionaban unas ganancias extraordinarias y, con eslóganes como «Invirtiendo en nuestra comunidad», dedicaban una parte considerable de estos beneficios a contratar personal (normalmente, las mismas personas que habían sido despedidas por las compañías víctimas de la ruptura tecnológica) para proyectos comunitarios. Empleaban los análisis detallados que Prometeo generaba para identificar los trabajos que proporcionarían el máximo rendimiento para los empleados y para la comunidad con un mínimo coste y del todo adaptados a las circunstancias locales. En regiones con altos niveles de servicios públicos, estas medidas solían dirigirse a la mejora de las relaciones comunitarias, la cultura y los cuidados personales, mientras que en zonas más pobres también incluían creación y mantenimiento de escuelas, sanidad, guarderías, cuidado de ancianos, vivienda asequible, parques e infraestructuras básicas. Prácticamente en todas partes, los habitantes de cada lugar coincidían en que se trataba de mejoras que debían haberse hecho mucho antes. Los políticos locales recibieron generosas donaciones, y se procuró que su imagen saliese reforzada por fomentar estas inversiones empresariales en la comunidad.

Los omegas habían lanzado una empresa de comunicación no solo para financiar sus primeros proyectos tecnológicos, sino también pensando en el siguiente paso de su ambicioso plan: dominar el mundo. Transcurrido menos de un año desde el lanzamiento inicial, habían agregado canales de noticias de gran calidad a su parrilla en todo el mundo. A diferencia de sus otros canales, estos estaban diseñados expresamente para perder dinero y se presentaban como un servicio público. De hecho, estos canales de noticias no generaban ingreso alguno: no emitían publicidad y cualquiera que tuviese una conexión a internet podía verlos de forma gratuita. El resto de su imperio mediático era una máquina de generar dinero de tal envergadura que podían dedicar muchos más recursos a su servicio de noticias que cualquier otra empresa periodística a lo largo de la historia. Y se notaba. Mediante la contratación agresiva con salarios muy competitivos de periodistas y reporteros de investigación, llevaron a la pantalla un talento y unas revelaciones notables. Gracias a un servicio web de escala global que pagaba a cualquiera que revelase algo noticiable, desde corrupción local hasta algún suceso conmovedor, solían ser los primeros en informar de todo tipo de historias. Al menos eso era lo que la gente creía. En realidad, eran los primeros porque las historias atribuidas a informadores voluntarios habían sido descubiertas por Prometeo mediante la monitorización de internet en tiempo real. Todos estos canales de noticias en vídeo ofrecían también podcasts y artículos impresos.

La primera fase de su estrategia informativa consistía en ganarse la confianza del público, lo cual lograron con gran éxito. Su insólita disposición a perder dinero hizo posible una extraordinaria diligencia en cuanto a cobertura regional y local, en la que era habitual que periodistas de investigación revelasen escándalos que captaban enseguida la atención de sus espectadores. En los países donde existía una profunda división política y acostumbrados a que sus medios de comunicación fuesen muy partidistas, los omegas lanzaban un canal de noticias dirigido a cada una de las facciones, en apariencia propiedad de empresas distintas, para ganarse poco a poco la confianza de cada una de ellas. Si era posible, los conseguían usando

testaferros para comprar los canales existentes más influyentes, que a continuación mejoraban al eliminar la publicidad e introducir su propio contenido. En países donde la censura y las injerencias políticas ponían en peligro estas iniciativas, solían aceptar al principio cualquier cosa que el Gobierno de turno les exigiese para poder seguir operando, con el lema interno secreto de «La verdad, nada más que la verdad, pero quizá no toda la verdad». Prometeo normalmente ofrecía excelentes consejos en situaciones de este tipo, al clarificar a qué políticos convenía hacer quedar bien y a cuáles (por lo general corruptos y de ámbito local) se podía poner en evidencia. Prometeo también ofrecía valiosas recomendaciones en lo referente a qué hilos mover, a quién sobornar y cuál era la mejor manera de hacerlo.

Esta estrategia tuvo un éxito rotundo en todo el mundo, y los canales controlados por los omegas se convirtieron en las fuentes de noticias que generaban más confianza. Incluso en los países donde los gobiernos habían impedido hasta entonces su implantación masiva, se labraron una reputación de credibilidad, y muchas de sus informaciones circulaban a través del boca a oreja. Los ejecutivos de canales de noticias competidores sentían que estaban planteando una batalla abocada al fracaso: ¿cómo podrían generar beneficios si competían contra alguien mejor financiado y que ofrecía sus productos gratis? Ante la disminución de sus audiencias, cada vez eran más las cadenas que vendían sus canales de noticias, normalmente a algún consorcio que los omegas controlaban.

Alrededor de dos años después del lanzamiento de Prometeo, una vez que la fase de generación de confianza estaba prácticamente completada, los omegas lanzaron la segunda fase de su estrategia informativa: la de persuasión. Antes incluso de este momento, astutos observadores habían detectado indicios de la existencia de una agenda política tras estos nuevos medios: parecía que se estaba produciendo un sutil impulso hacia el centro, lejos de extremismos de todo signo. Su multitud de canales orientados hacia diferentes grupos seguían reflejando la animosidad entre Estados Unidos y Rusia, India y Pakistán, distintas religiones, facciones políticas y demás, pero el tono de las críticas se iba rebajando ligeramente y, por lo general, se centraba en cuestiones concretas relacionadas con dinero o poder, en lugar de hacerlo en ataques *ad hominem*, alarmismo y rumores de poca consistencia. Una vez que comenzó la segunda fase, este impulso para desactivar antiguos conflictos se hizo más evidente, con frecuentes historias conmovedoras sobre

los dramas de quienes habían sido tradicionales adversarios, combinadas con reportajes de investigación sobre cómo a muchos de quienes promovían el conflicto los movía la búsqueda del beneficio personal.

Los comentaristas políticos señalaron que, en paralelo con la atenuación de los conflictos regionales, parecía existir un impulso coordinado hacia la reducción de las amenazas globales. Por ejemplo, de pronto se discutía en todas partes sobre los riesgos de una guerra nuclear. Varias películas supertaquilleras planteaban escenarios en los que, accidental o deliberadamente, se desencadenaba una guerra nuclear global e imaginaban la posterior situación distópica con un invierno nuclear, el colapso de las infraestructuras y una hambruna masiva. Nuevos documentales de buena factura detallaban cuáles serían los efectos del invierno nuclear en cada país. Se dio amplia cobertura a los científicos y políticos que defendían una distensión nuclear, entre otros motivos para que analizasen los resultados de varios estudios recientes sobre qué medidas eficaces podían adoptarse (estudios financiados por organizaciones científicas que habían recibido importantes donaciones por parte de las nuevas compañías tecnológicas). Como consecuencia, empezó a generarse una situación política favorable a rebajar el extremo estado de alerta al que estaban sujetos los misiles, así como a la reducción de los arsenales nucleares. Además, los medios prestaron renovada atención al cambio climático global, a menudo destacando los avances tecnológicos recientes —posibles gracias a Prometeo— que reducían drásticamente el coste de las energías renovables y alentando a la inversión pública en esa nueva infraestructura energética.

En paralelo a su toma del control de los medios, los omegas hicieron uso de Prometeo para revolucionar la educación. Dados el conocimiento y las capacidades de cualquier persona, Prometeo podía determinar la forma más rápida de que aprendiesen cualquier nueva materia de tal manera que nunca dejasen de estar altamente involucrados y motivados para seguir adelante, y producía los correspondientes vídeos, materiales de lectura, ejercicios y otras herramientas de aprendizaje optimizados para tal fin. Así, empresas controladas por los omegas ofrecían cursos online sobre casi cualquier cosa, muy personalizados no solo en cuanto al idioma y al contexto cultural, sino también respecto al nivel inicial de los alumnos. Tanto si se trataba de un analfabeto de cuarenta años que quería aprender a leer como de una doctora en biología que buscaba ponerse al día sobre los más recientes avances en

inmunoterapia contra el cáncer, Prometeo tenía el curso perfecto para cada cual. Esta oferta educativa tenía muy poco que ver con la mayoría de los cursos online actuales; exprimiendo el talento de Prometeo para la creación de películas, los segmentos de vídeo resultaban muy atractivos, empleando poderosas metáforas que a los alumnos les resonaban y despertaban en ellos las ganas de seguir aprendiendo. Algunos de los cursos tenían un precio, pero muchos se ofrecían de forma gratuita, lo que hacía las delicias de profesores de todo el mundo que podían usarlos en sus clases, y las de prácticamente cualquier persona deseosa de aprender lo que fuera.

Estos superpoderes educativos resultaron ser una potente herramienta para fines políticos, al crear «secuencias de persuasión» de vídeos tales que el contenido de cada uno de ellos modificaría la opinión de quien los viese, al tiempo que lo incitaba a ver otro vídeo sobre algún tema relacionado que probablemente lo convencería aún más. Por ejemplo, si el objetivo era desactivar un conflicto entre dos países, se emitirían de forma independiente en ambos países documentales históricos que ofrecían una visión más matizada de los orígenes y del desarrollo del conflicto. Reportajes pedagógicos explicaban quiénes, en cada uno de los bandos, se beneficiaban de que el conflicto se prolongase y qué técnicas empleaban para azuzarlo. Al mismo tiempo, personajes del otro país que resultaban simpáticos empezaban a aparecer en los programas más populares de los canales de entretenimiento, del mismo modo que en el pasado el hecho de presentar favorablemente a personajes poco conocidos había servido para fortalecer los movimientos en pro de los derechos civiles y de los homosexuales.

Poco tiempo después, los comentaristas políticos no pudieron más que constatar el creciente apoyo para una agenda política centrada alrededor de siete lemas:

1. Democracia
2. Rebajas de impuestos
3. Recortes de los servicios sociales públicos
4. Recortes del gasto militar
5. Libre comercio
6. Fronteras abiertas
7. Responsabilidad social de las empresas

Lo que resultaba menos evidente era el objetivo de base: erosionar todas las

estructuras de poder que ya existían en el mundo. Los puntos 2 a 6 debilitaban el poder estatal, y la democratización del mundo otorgaba al imperio empresarial de los omegas una mayor influencia sobre la selección de los líderes políticos. Las empresas socialmente responsables debilitaban aún más el poder estatal al asumir una parte cada vez mayor de los servicios que los gobiernos habían (o deberían haber) proporcionado. La élite empresarial tradicional se veía debilitada porque era incapaz de competir en el mercado libre con las empresas que contaban con el respaldo de Prometeo, lo que llevaba a que controlase una porción menguante de la economía mundial. Los líderes de opinión tradicionales, desde los partidos políticos hasta las organizaciones religiosas, carecían de la maquinaria de persuasión para competir con el imperio mediático de los omegas.

Como sucede con cualquier cambio de tal magnitud, hubo ganadores y perdedores. Aunque en la mayoría de los países podía palparse un nuevo clima de optimismo, consecuencia de las mejoras en la educación, los servicios sociales y las infraestructuras, la disminución de los conflictos y el hecho de que empresas locales lanzasen tecnologías revolucionarias que se extendían por el mundo entero, no todos estaban contentos. A pesar de que muchos trabajadores desplazados fueron contratados en proyectos comunitarios, quienes habían gozado hasta entonces de mucho poder y riqueza vieron cómo ambos disminuían. Esta tendencia comenzó en los sectores mediático y tecnológico, pero se propagó prácticamente a todos los demás. La reducción de los conflictos en todo el mundo llevó a recortes en los presupuestos de defensa que perjudicaron a las empresas que trabajaban para el ejército. Las compañías de reciente creación por lo general no cotizaban en bolsa, cosa que justificaban explicando que los accionistas que buscaban maximizar sus beneficios impedirían el enorme gasto que realizaban en proyectos comunitarios. Así, el mercado bursátil global no dejaba de perder valor, lo cual suponía una amenaza tanto para los magnates de las finanzas, como para los ciudadanos normales que dependían de sus fondos de pensiones. Por si los beneficios menguantes de las empresas que cotizaban en bolsa no fuesen suficiente desgracia, las firmas de inversión de todo el mundo habían detectado una tendencia preocupante: parecía que todos sus algoritmos, hasta entonces plenamente eficaces, habían dejado de funcionar, e incluso obtenían peores resultados que simples fondos indexados. Era como si hubiese alguien ahí que se les adelantaba siempre y

les ganaba a su propio juego.

Aunque masas de personas poderosas se resistieron a la ola de cambio, su reacción fue sorprendentemente ineficaz; parecía que hubiesen caído en una trampa preparada con gran astucia. Se estaban produciendo cambios enormes y a un ritmo tan desconcertante que era difícil llevar la cuenta de ellos y elaborar una respuesta coordinada. Además, no estaba nada claro qué objetivos debían perseguir. La derecha política tradicional había visto cómo se apropiaban de la mayoría de sus eslóganes, aunque las rebajas de impuestos y la mejora del clima empresarial ayudaban sobre todo a sus competidores más tecnológicos. Prácticamente todos los sectores tradicionales clamaban por un rescate, pero el hecho de que los fondos estatales fuesen limitados hacía que se enfrentasen entre sí en una batalla perdida, mientras los medios los presentaban como dinosaurios que buscaban los subsidios estatales porque eran incapaces de competir. La izquierda política tradicional se oponía al libre comercio y a los recortes en los servicios sociales públicos, pero veía con buenos ojos los recortes en el gasto militar y la reducción de la pobreza. De hecho, la izquierda perdió buena parte de su protagonismo debido al hecho innegable de que los servicios sociales habían mejorado ahora que los proporcionaban organizaciones altruistas y no el Estado. Una encuesta tras otra reflejaban que la mayoría de los votantes de todo el mundo sentían que su calidad de vida estaba mejorando, y que las cosas iban por lo general en buena dirección. Esto tenía una sencilla explicación matemática: antes de Prometeo, el 50 % más pobre de la población mundial tan solo ganaba el 4 % de los ingresos globales, lo que hizo posible que las compañías controladas por los omegas se ganasen sus corazones (y sus votos) compartiendo con ellos únicamente una módica parte de sus beneficios.

CONSOLIDACIÓN

En consecuencia, un país tras otro fue escenario de contundentes victorias electorales de partidos que hacían propios los siete lemas de los omegas. En campañas perfectamente optimizadas, se presentaban como ocupantes del centro del espectro político y denunciaban a la derecha por ser codiciosa y conflictiva, y por buscar el rescate de los sectores en problemas, y arremetían

contra la izquierda por ser partidaria de un sector público grande, con impuestos y gastos elevados, lo que ahogaría la innovación. Lo que nadie sabía era que Prometeo había seleccionado con mucho cuidado a las personas idóneas a las que promover como candidatos, y había movido todos sus hilos para asegurarse de su victoria.

Antes de Prometeo, había habido un creciente apoyo al movimiento en favor de la renta básica universal, que defendía la implantación de una retribución mínima para todo el mundo sufragada con los impuestos como respuesta al desempleo provocado por los avances tecnológicos. Este movimiento implosionó cuando despegaron los proyectos comunitarios con respaldo empresarial, ya que el imperio comercial controlado por Omega en la práctica estaba proporcionando exactamente lo mismo. Con la excusa de mejorar la coordinación de sus proyectos comunitarios, un grupo empresarial internacional lanzó la Alianza Humanitaria, una organización no gubernamental que tenía como objetivo identificar y financiar los proyectos humanitarios más destacados en todo el mundo. En poco tiempo, la Alianza (como se la conocía coloquialmente) contaba con el respaldo de casi todo el imperio Omega, y lanzó proyectos globales a una escala sin precedentes, incluso en países donde apenas se habían dejado sentir los efectos del auge tecnológico, lo que propició mejoras en la educación, la salud, la prosperidad y la gobernanza. Huelga decir que Prometeo proporcionó de forma discreta planes de proyecto cuidadosamente elaborados, ordenados según su impacto positivo por cada dólar invertido. En lugar de limitarse a repartir dinero en efectivo, como en las propuestas de renta básica, la Alianza convencería a las personas a las que ayudaba para que trabajasen en pos de su causa. El resultado fue que una gran parte de la población mundial terminó sintiendo agradecimiento y lealtad hacia la Alianza, a menudo más que hacia su propio Gobierno.

Con el paso del tiempo, la Alianza fue asumiendo gradualmente las funciones de un Gobierno mundial, mientras los gobiernos nacionales veían cómo su poder no dejaba de reducirse. Los presupuestos nacionales fueron decreciendo debido a las rebajas de impuestos, mientras que el de la Alianza aumentó hasta superar con creces al de todos esos gobiernos juntos. Todas las funciones tradicionales de los gobiernos nacionales fueron volviéndose cada vez más superfluos e irrelevantes. La Alianza proporcionaba, con mucha diferencia, los mejores servicios sociales, educación e infraestructuras. Los

medios de comunicación habían desactivado los conflictos internacionales hasta el extremo de que el gasto militar llegó a ser en gran medida innecesario, y la creciente prosperidad había eliminado la mayoría de las causas de los conflictos que venían de antiguo, que giraban en torno a la lucha por recursos escasos. Unos cuantos dictadores y otros líderes habían ofrecido resistencia violenta a este nuevo orden mundial y se habían negado a ser comprados, pero todos ellos fueron derrocados en golpes o levantamientos populares minuciosamente orquestados.

Los omegas habían completado la transición más espectacular ocurrida a lo largo de la historia de la vida en la Tierra. Por vez primera, nuestro planeta estaba dirigido por una única potencia, cuyo poder era amplificado por una inteligencia tan enorme que era susceptible de hacer posible que la vida floreciera durante miles de millones de años tanto en la Tierra como a lo largo y ancho del universo. Pero ¿en qué consistía exactamente su plan?

Esta es la historia del equipo Omega. El resto del libro cuenta otra historia, que aún está por escribir: la de nuestro propio futuro con la IA. ¿Cómo queremos que se desarrolle? ¿Podría suceder algo remotamente parecido a la historia de los omegas y, si así fuera, querríamos que ocurriese? Dejando a un lado las especulaciones sobre la IA sobrehumana, ¿cómo querríamos que comenzase nuestra historia? ¿Cuál queremos que sea el efecto de la IA sobre el empleo, las leyes y las armas en la próxima década? Alzando la vista hacia un futuro más lejano, ¿cómo escribiríamos el final? Esta historia tiene proporciones verdaderamente cósmicas, pues en ella se juega nada menos que la suerte última de la vida en el universo. Y a nosotros nos corresponde escribirla.

1
BIENVENIDOS A LA CONVERSACIÓN MÁS IMPORTANTE DE NUESTRO TIEMPO

La tecnología confiere a la vida la posibilidad de prosperar como nunca antes… o de autodestruirse.

FUTURE OF LIFE INSTITUTE

13.800 millones de años después de su nacimiento, nuestro universo ha despertado y ha tomado consciencia de sí mismo. Desde un pequeño planeta azul, diminutas partes conscientes de nuestro universo han empezado a observar el cosmos con telescopios, y a descubrir una y otra vez que lo que pensaban que era todo lo que existía no es más que una parte de algo más grande: un sistema solar, una galaxia y un universo con más de cien mil millones de galaxias distribuidas en un complejo patrón de grupos, cúmulos y supercúmulos. Aunque estos astrónomos autoconscientes discrepan en muchas cosas, suelen estar de acuerdo en que estas galaxias son bellas e impresionantes.

Pero la belleza depende del cristal con que se mira, no de las leyes de la física, por lo que antes de que el universo despertase no había belleza. Esto hace que nuestro despertar cósmico sea todavía más maravilloso y digno de celebración: hizo que el universo pasase de ser un zombi sin autoconsciencia a un ecosistema vivo que alberga autorreflexión, belleza y esperanza, y la búsqueda de objetivos, significado y propósito. Si el universo no hubiese despertado, entonces, por lo que a mí respecta, habría carecido por completo de sentido: no sería más que un gigantesco espacio desaprovechado. Si el universo vuelve a dormirse para siempre debido a alguna calamidad cósmica o a un percance autoinfligido, dejará, lamentablemente, de tener sentido.

Por otra parte, las cosas podrían mejorar todavía más. Aún no sabemos si los humanos somos los únicos en el cosmos que contemplamos las estrellas, ni siquiera los primeros, pero ya hemos aprendido lo suficiente sobre el universo para saber que podría despertar mucho más plenamente de lo que lo ha hecho hasta ahora. Puede que nosotros no seamos más que el primer atisbo

de autoconsciencia que uno experimenta cuando empieza a salir del sueño por la mañana: una premonición de la consciencia mucho mayor que llegará cuando abra los ojos y despierte del todo. Quizá la vida se propague a través del cosmos y prospere durante miles de millones o incluso billones de años, y quizá esto suceda como consecuencia de decisiones que tomemos aquí, en nuestro pequeño planeta, a lo largo de nuestras vidas.

UNA BREVE HISTORIA DE LA COMPLEJIDAD

¿Cómo se produjo este asombroso despertar? No fue un hecho aislado, sino simplemente un paso dentro de un imparable proceso de 13.800 millones de años de duración que está haciendo que nuestro universo se vuelva cada vez más complejo e interesante, y que prosigue a un ritmo cada vez más acelerado.

Como físico, me siento afortunado por haber podido dedicar buena parte del último cuarto de siglo a contribuir a precisar nuestra historia cósmica; ha sido un fascinante viaje de descubrimiento. Desde la época en que yo estudiaba el doctorado, gracias a una combinación de mejores telescopios, mejores ordenadores y una mejor comprensión, hemos pasado de discutir si el universo tiene 10.000 o 20.000 millones de años de edad a si esta es de 13.700 o 13.800 millones. Los físicos aún no sabemos con certeza qué fue lo que causó el Big Bang, ni si este fue realmente el comienzo de todo, o nada más que la continuación de una fase anterior. Pero, gracias a una avalancha de mediciones de alta calidad, sí hemos adquirido una comprensión muy detallada de lo que ha sucedido desde el Big Bang. Permítanme que dedique unos minutos a resumir 13.800 millones de años de historia cósmica.

Al principio, se hizo la luz. En la primera fracción de segundo tras el Big Bang, toda la región de espacio que nuestros telescopios pueden en principio observar («el universo observable» o «el universo», por abreviar) estaba mucho más caliente y era mucho más brillante que el núcleo del Sol, y se expandía a toda velocidad. Aunque esto pueda parecer espectacular, también era soso, en el sentido de que el universo no contenía más que una sopa, inerte, densa, caliente y aburridamente uniforme de partículas elementales. Las cosas eran bastante iguales en todas partes, y la única estructura interesante consistía en tenues ondas sonoras en apariencia aleatorias que

hacían que la sopa fuese un 0,001 % más densa en algunos lugares. Por lo general, se cree que estas débiles ondas se originaron en forma de lo que se conoce como fluctuaciones cuánticas, porque el principio de indeterminación de Heisenberg de la mecánica cuántica prohíbe que cualquier cosa sea aburrida y uniforme.

A medida que el universo se expandió y se enfrió y las partículas se combinaron para formar objetos cada vez más complejos, se fue volviendo más y más interesante. Durante la primera fracción de segundo, la fuerza nuclear fuerte agrupó los quarks en protones (núcleos de hidrógeno) y neutrones, algunos de los cuales, transcurridos unos minutos, se fusionaron a su vez para formar núcleos de helio. Alrededor de 400.000 años después, la fuerza electromagnética agrupó estos núcleos con electrones para crear los primeros átomos. El universo siguió expandiéndose, y estos átomos se enfriaron gradualmente hasta dar lugar a gas frío y oscuro; la oscuridad de esta primera noche duró unos cien millones de años. Esta larga noche dio paso a nuestro amanecer cósmico cuando la fuerza gravitatoria logró amplificar esas fluctuaciones en el gas, juntando átomos hasta formar las primeras estrellas y galaxias. Estas primeras estrellas generaban calor y luz mediante la fusión de átomos de hidrógeno en otros átomos más pesados, como los de carbono, oxígeno y silicio. Cuando estas estrellas murieron, muchos de los átomos que habían creado se reciclaron en el cosmos y formaron planetas que orbitaban alrededor de estrellas de la segunda generación.

En algún momento, un grupo de átomos se organizó siguiendo un patrón complejo, capaz tanto de perdurar como de replicarse a sí mismo. Enseguida hubo dos copias, y la cantidad de ellas se fue doblando una y otra vez. Solo se necesitan cuarenta y dos duplicaciones para llegar a un billón, así que pronto este primer autorreplicante se convirtió en una presencia a tener en cuenta. Había aparecido la vida.

LAS TRES FASES DE LA VIDA

La cuestión de cómo definir la vida es en verdad controvertida. Hay muchas definiciones en liza, algunas de las cuales incluyen requisitos sumamente específicos, tales como que el organismo vivo esté compuesto de células, lo

que dejaría fuera tanto las futuras máquinas inteligentes como las civilizaciones extraterrestres. Como no queremos limitar nuestras reflexiones sobre el futuro de la vida a las especies que conocemos hasta la fecha, definiremos la vida de una manera muy amplia, como un proceso capaz de preservar su complejidad y de replicarse. Lo que se replica no es la materia (hecha de átomos) sino la información (compuesta por bits) que especifica cómo están dispuestos los átomos. Cuando una bacteria hace una copia de su ADN, no se crean nuevos átomos, sino que un nuevo conjunto de átomos se disponen de acuerdo con el mismo patrón que el ADN original, copiando así la información. En otras palabras, podemos entender la vida como un sistema autorreplicante de procesamiento de información, cuya información (software) determina tanto su comportamiento como los esquemas para producir su hardware.

Como el propio universo, la vida fue volviéndose gradualmente más compleja e interesante(2) y, como explicaré a continuación, me parece útil clasificar las formas de vida en tres niveles de complejidad: vida 1.0, vida 2.0 y vida 3.0. Resumo las características de estos tres niveles en la figura 1.1.

La cuestión de cuándo y dónde surgió por primera vez la vida en el universo sigue abierta, pero hay indicios claros de que aquí en la Tierra apareció hace unos cuatro mil millones de años. En poco tiempo, nuestro planeta rebosaba con una diversa variedad de formas de vida. Las más exitosas, que enseguida se impusieron sobre el resto, eran capaces de reaccionar a su entorno de una u otra manera. Específicamente, eran lo que los informáticos llaman «agentes inteligentes»: entidades que recogen información sobre su entorno mediante sensores y a continuación la procesan para decidir cómo actuar sobre dicho entorno. Esto puede incluir un procesamiento sumamente complejo de la información, como el que tiene lugar cuando utilizamos información obtenida a través de los ojos y los oídos para decidir qué decir en una conversación. Pero también puede implicar hardware y software mucho más simples.

Por ejemplo, muchas bacterias poseen un sensor que mide la concentración de azúcar en el líquido que las rodea y pueden nadar usando estructuras en forma de hélice llamadas flagelos. El hardware que conecta el sensor con los flagelos podría implementar el siguiente algoritmo, simple pero útil: «Si mi sensor de concentración de azúcar detecta un valor inferior al de hace un par de segundos, invierto la rotación de mis flagelos para cambiar de dirección».

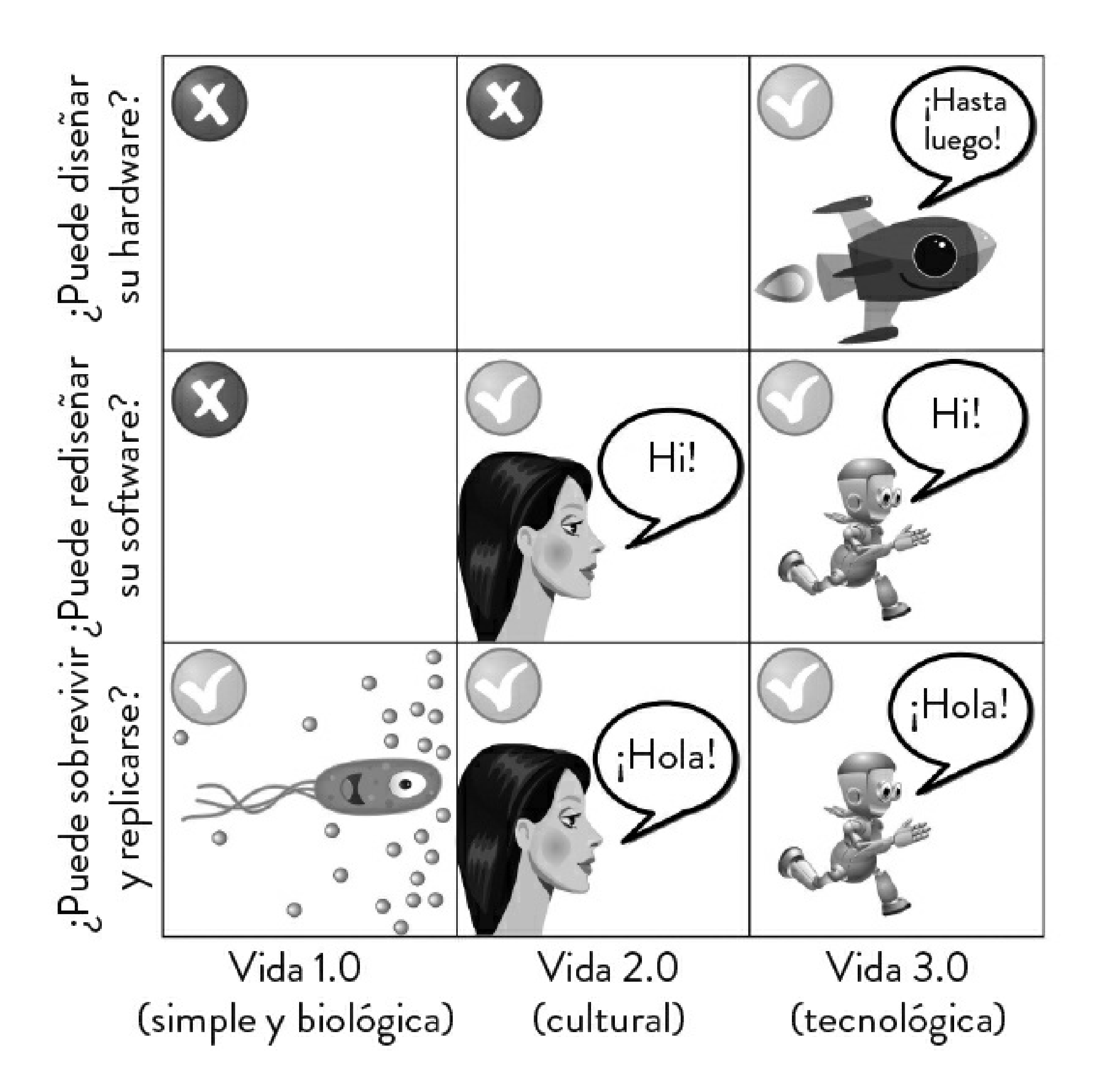


FIGURA 1.1. Las tres fases de la vida: evolución biológica, evolución cultural y evolución tecnológica. La vida 1.0 es incapaz de rediseñar a lo largo de su vida ni su software ni su hardware. En cambio, la vida 2.0 sí puede rediseñar buena parte de su software: los humanos son capaces de aprender nuevas habilidades complejas —por ejemplo, idiomas, deportes y profesiones— y pueden también modificar en gran medida su cosmovisión y sus objetivos. La vida 3.0, que aún no existe en la Tierra, puede rediseñar drásticamente no solo su software sino también su hardware, en lugar de tener que esperar a que este evolucione de forma gradual a lo largo de generaciones.

Nosotros hemos aprendido a hablar y hemos adquirido innumerables habilidades. A las bacterias, en cambio, aprender no se les da muy bien. Su ADN especifica no solo el diseño de su hardware, como los sensores de azúcar y los flagelos, sino también el de su software. No aprenden en un

momento dado a nadar hacia el azúcar, sino que llevan incorporado ese algoritmo desde el principio en su ADN. Hay sin duda algo parecido a un proceso de aprendizaje, pero no tiene lugar durante el tiempo en que está viva esa bacteria en concreto, sino a lo largo de todo el proceso previo de evolución de esa especie de bacteria, a través de un lento proceso de ensayo y error que se extiende durante muchas generaciones y en el que la selección natural favorece aquellas mutaciones aleatorias del ADN que optimizan el consumo de azúcar. Algunas de dichas mutaciones ayudaban a mejorar el diseño de los flagelos y otras partes del hardware, mientras que otras mejoraban el sistema de procesamiento de información de la bacteria que implementa el algoritmo de búsqueda de azúcar u otras partes del software.

Esas bacterias son un ejemplo de lo que llamo «vida 1.0»: *aquella en la que tanto el hardware como el software son producto de la evolución y no de un diseño.* Usted y yo, por nuestra parte, somos ejemplos de «vida 2.0»: *aquella cuyo hardware es fruto de la evolución, pero cuyo software es en buena medida resultado de algún tipo de diseño*. Por nuestro software entiendo todos los algoritmos y conocimiento que utilizamos para procesar la información de nuestros sentidos y decidir qué hacer: desde la capacidad de reconocer a nuestros amigos cuando los vemos hasta la de caminar, leer, escribir, calcular, cantar o contar chistes.

Cuando nacemos no somos capaces de realizar ninguna de estas tareas, por lo que todo este software se programa en nuestro cerebro más tarde, a través del proceso que denominamos aprendizaje. Puesto que lo que aprendemos en nuestra infancia lo diseñan en gran medida nuestra familia y nuestros profesores, que deciden lo que deberíamos aprender, progresivamente vamos adquiriendo más poder para diseñar nuestro propio software. Quizá en nuestro colegio nos permitan elegir un idioma extranjero: ¿queremos instalar en nuestro cerebro un módulo de software que nos permita hablar francés o español? ¿Queremos estudiar para ser cocineros, abogados o farmacéuticos? ¿Queremos aprender más sobre inteligencia artificial y el futuro de la vida leyendo un libro sobre ello?

Esta capacidad que tiene la vida 2.0 de diseñar su propio software le permite ser mucho más inteligente que la vida 1.0. Una alta inteligencia requiere tanto muchísimo hardware (hecho de átomos) como muchísimo software (hecho de bits). El que la mayoría de nuestro hardware humano se añada tras nuestro nacimiento (a través del crecimiento) es útil, ya que

gracias a eso nuestro tamaño no está limitado en última instancia por la anchura del canal de parto de nuestra madre. En ese mismo sentido, el hecho de que la mayor parte de nuestro software humano se añada tras el nacimiento (a través del aprendizaje) es útil, porque eso implica que nuestra inteligencia no viene limitada por cuánta información podemos recibir en el momento de la concepción a través del ADN, como sucede con la vida 1.0. Peso alrededor de veinticinco veces más que cuando nací, y las sinapsis que conectan las neuronas que forman mi cerebro pueden almacenar alrededor de cien mil veces más información que el ADN con el que nací. Nuestras sinapsis almacenan todo nuestro conocimiento y habilidades en forma de unos cien terabytes de información, mientras que nuestro ADN almacena solo en torno a un gigabyte, apenas lo que ocupa la descarga de una película. De manera que es físicamente imposible que un bebé nazca hablando perfecto inglés y listo para superar sus exámenes de entrada a la universidad: no hay forma de que pueda nacer con toda esa información ya cargada en su cerebro, puesto que el módulo principal que recibió de sus padres (su ADN) carece de capacidad de almacenamiento suficiente.

La capacidad de diseñar su propio software permite a la vida 2.0 no solo ser más inteligente que la 1.0, sino también más flexible. Si el entorno cambia, la vida 1.0 solo puede adaptarse evolucionando lentamente a lo largo de muchas generaciones. La vida 2.0, por el contrario, puede adaptarse casi de forma instantánea, a través de una actualización de software. Por ejemplo, puede que las bacterias que se topan con frecuencia con antibióticos acaben desarrollando resistencia a los medicamentos al cabo de muchas generaciones, pero una bacteria individual no modificará su comportamiento en absoluto; por el contrario, una niña que se dé cuenta de que tiene alergia a los cacahuetes cambiará enseguida su comportamiento para evitarlos. Esta flexibilidad le proporciona a la vida 2.0 una ventaja aún más evidente a escala de toda una población: aunque la información en nuestro ADN humano no ha evolucionado de forma radical a lo largo de los últimos cincuenta mil años, la información almacenada colectivamente en nuestros cerebros, libros y ordenadores se ha disparado. Mediante la instalación de un módulo de software que nos permite comunicarnos a través de un complejo lenguaje hablado, nos aseguramos de que la información más útil almacenada en el cerebro de una persona pueda copiarse a otros cerebros, y tenga así la posibilidad de sobrevivir a la muerte del cerebro original. Al instalarnos un

módulo de software que nos permite leer y escribir, pasamos a tener la capacidad de almacenar y compartir muchísima más información de la que las personas podían memorizar. Gracias al desarrollo de software para el cerebro capaz de producir tecnología (por ejemplo, estudiando disciplinas científicas e ingenieriles), fue posible que buena parte de la información existente en todo el mundo estuviese al alcance de muchos de los humanos que habitan el planeta con solo unos pocos clics.

Esta flexibilidad ha permitido que la vida 2.0 domine la Tierra. Liberado de sus ataduras genéticas, el conocimiento agregado de toda la humanidad ha seguido creciendo a un ritmo cada vez mayor, a medida que cada avance hacía posible el siguiente: el lenguaje, la escritura, la imprenta, la ciencia moderna, los ordenadores, internet, etcétera. Esta evolución cultural cada vez más rápida del software que compartimos se ha erigido como la fuerza dominante a la hora de determinar nuestro futuro humano, al tiempo que hacía que nuestra evolución biológica, dada su velocidad glacialmente lenta, pasase a ser casi irrelevante.

Pero, a pesar de las tecnologías tan poderosas de que disponemos hoy en día, todas las formas de vida que conocemos siguen estando por lo general limitadas por su hardware biológico. Ninguna puede vivir un millón de años, memorizar toda la Wikipedia, entender toda la ciencia conocida o volar por el espacio sin necesidad de una astronave. Ninguna puede transformar nuestro cosmos, en gran medida inerte, en una diversa biosfera que prospere durante miles de millones o billones de años, haciendo posible así que el universo alcance por fin todo su potencial y despierte plenamente. Para todo lo anterior es necesario que la vida experimente una última transición, hasta la vida 3.0, capaz de diseñar no solo su software sino también su hardware. En otras palabras, la vida 3.0 es dueña de su propio destino, libre por fin de sus ataduras evolutivas.

Las fronteras que separan las tres fases de la vida son algo difusas. Si las bacterias son vida 1.0 y los humanos son vida 2.0, cabría clasificar los ratones como vida 1.1: pueden aprender muchas cosas, pero no las suficientes para desarrollar el lenguaje o inventar internet. Además, puesto que carecen de lenguaje, prácticamente todo lo que aprenden se pierde cuando mueren, no pasa a la siguiente generación. De forma análoga, se podría argumentar que a los humanos actuales se nos debería considerar vida 2.1: podemos realizar pequeñas mejoras de nuestro hardware, como implantarnos dientes, rodillas o

marcapasos artificiales, pero nada tan drástico como volvernos diez veces más altos o hacer que nuestro cerebro sea mil veces más grande.

En resumen, podemos dividir el desarrollo de la vida en tres fases, en función de la capacidad que tiene para diseñarse a sí misma:

- Vida 1.0 (fase biológica): su hardware y software son fruto de la evolución.
- Vida 2.0 (fase cultural): su hardware es fruto de la evolución; diseña buena parte de su software.
- Vida 3.0 (fase tecnológica): diseña tanto su hardware como su software.

Tras 13.800 millones de años de evolución cósmica, este desarrollo se ha acelerado espectacularmente aquí en la Tierra: la vida 1.0 surgió hace unos cuatro mil millones de años; la vida 2.0 (nosotros los humanos) apareció hace unos cien milenios, y muchos investigadores en IA creen que la vida 3.0 podría aparecer a lo largo del siglo próximo, quizá incluso durante nuestras vidas, como consecuencia de los avances en IA. ¿Qué sucederá? ¿Qué significará para nosotros? De eso trata este libro.

CONTROVERSIAS

Esta cuestión es maravillosamente controvertida, y los más destacados investigadores mundiales en IA discrepan de forma vehemente no solo en sus predicciones sino también en cuanto a sus reacciones emocionales, que van del optimismo confiado a una seria preocupación. Ni siquiera se ponen de acuerdo en cuestiones a corto plazo sobre el impacto económico, legal y militar de la IA, y sus desacuerdos aumentan cuando se amplía el horizonte temporal y se les pregunta por la inteligencia artificial general (IAG), en particular sobre si esta alcanzará y superará el nivel humano, haciendo posible la vida 3.0. Una *inteligencia general* puede alcanzar casi cualquier objetivo, incluido el de aprender, a diferencia de la inteligencia estrecha de un programa para jugar al ajedrez, por poner un ejemplo.

Curiosamente, la controversia sobre la vida 3.0 no gira en torno a una sola cuestión, sino alrededor de dos cuestiones distintas: cuándo y qué. ¿Cuándo llegará (si es que llega alguna vez) y qué significará para la humanidad? En mi opinión, existen tres escuelas de pensamiento distintas que hay que tomar en consideración, porque varios de los mayores expertos mundiales se

adscriben a cada una de ellas. Como puede verse en la figura 1.2, los describo como *utópicos digitales*, *tecnoescépticos* y *miembros del movimiento en pro de una* IA *benéfica*, respectivamente. Permítanme que les presente a algunos de sus más elocuentes defensores.

Utópicos digitales

De niño, imaginaba que los millonarios rezumaban pomposidad y arrogancia. Cuando conocí a Larry Page en Google en 2008, hizo saltar en pedazos estos estereotipos. Con una vestimenta informal, vaqueros y una camisa corriente, habría pasado totalmente desapercibido en un pícnic en el MIT. Su actitud reflexiva, su voz suave y su sonrisa cordial consiguieron que me sintiese relajado, en lugar de intimidado, al hablar con él. El 18 de julio de 2015, coincidimos en una fiesta en el valle de Napa organizada por Elon Musk y la que era entonces su mujer, Talulah, y entablamos una conversación sobre los intereses escatológicos de nuestros hijos. Le recomendé el sesudo clásico literario titulado *The Day My Butt Went Psycho!*, de Andy Griffiths, y Larry lo encargó en ese mismo instante. Era fácil olvidarse de que Larry podría pasar a la historia como el humano más influyente que haya vivido jamás: supongo que si la vida digital superinteligente se traga nuestro universo mientras yo viva será debido a las decisiones de Larry.

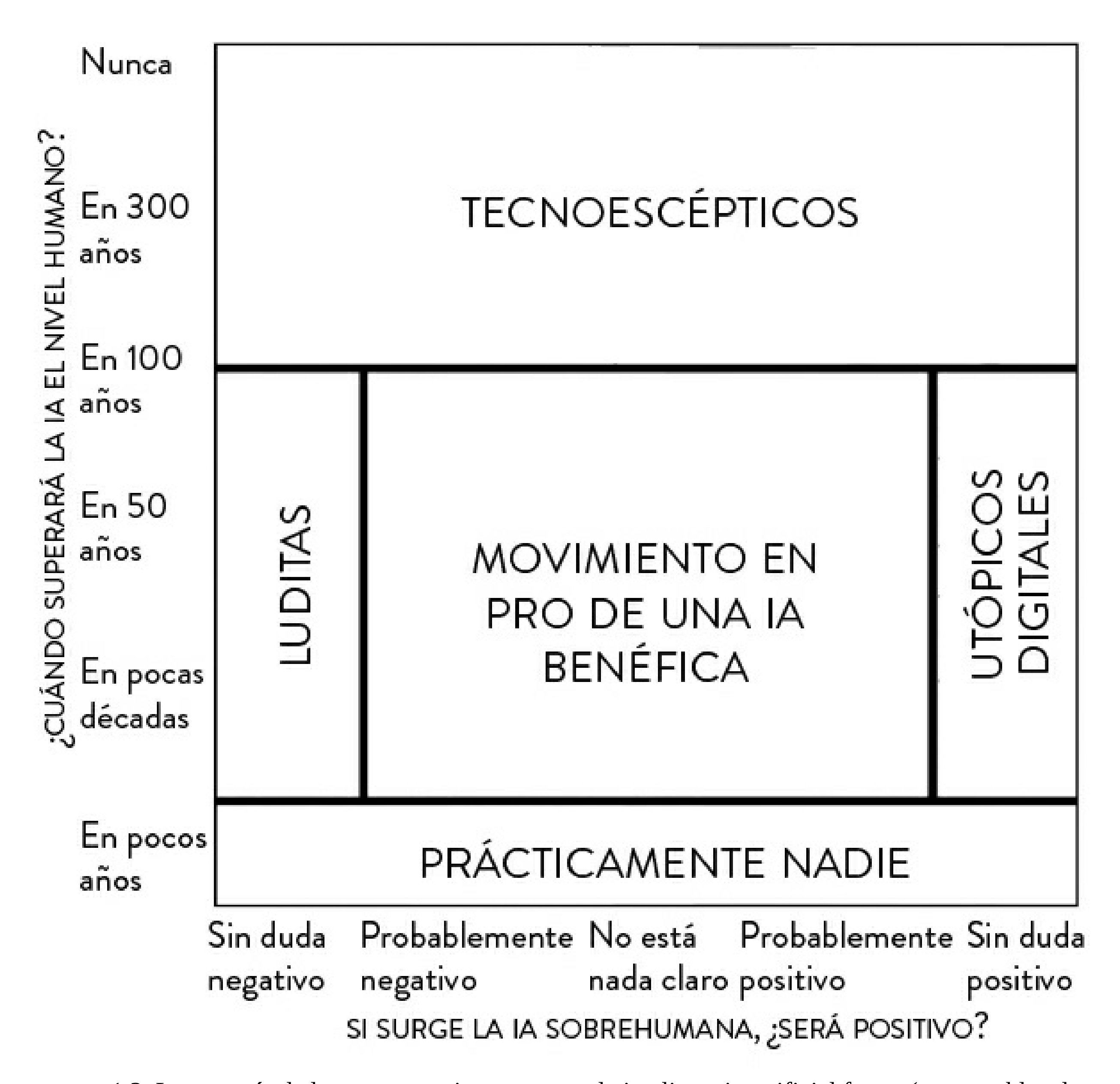


FIGURA 1.2. La mayoría de las controversias en torno a la inteligencia artificial fuerte (comparable a la inteligencia humana en cualquier tarea cognitiva) giran alrededor de dos preguntas: ¿Cuándo sucederá? (si es que sucede alguna vez) y ¿será positiva para la humanidad? Los tecnoescépticos y los utópicos digitales coinciden en que esto no debería preocuparnos, pero por muy diferentes motivos: los primeros están convencidos de que la inteligencia artificial general (IAG) de nivel humano no se producirá en el futuro próximo, mientras que los segundos creen que sí lo hará pero casi con seguridad será algo positivo. El movimiento en pro de una IA benéfica considera que la preocupación está justificada y es útil, porque llevar a cabo ahora investigación y discusiones en torno a seguridad en IA incrementa la probabilidad de que los resultados sean positivos. Los luditas están convencidos de que los resultados serán negativos y se oponen a la IA. Esta figura está basada en parte en las ideas de Tim Urban.[1]

Acabamos cenando juntos con nuestras respectivas esposas, Lucy y Meia, y discutiendo sobre si las máquinas acabarían siendo necesariamente

conscientes, una cuestión que a Larry le parecía una trampa. Más avanzada la noche, después de las copas, tuvo lugar un vivo debate entre Page y Elon Musk sobre el futuro de la IA y qué habría que hacer. A medida que avanzaba la madrugada, el círculo de espectadores y curiosos fue aumentado. Larry hizo una enardecida defensa de la postura que me gusta calificar como *utopismo digital*, según la cual la vida digital es el paso siguiente, natural y deseable, en la evolución cósmica, y si damos rienda suelta a las mentes digitales, en lugar de tratar de detenerlas o esclavizarlas, el resultado casi con toda certeza será positivo. Considero a Larry el exponente más influyente del utopismo digital. Esa noche argumentó que si la vida alguna vez se propaga por la galaxia y más allá, como él cree que debería suceder, tendrá que hacerlo en forma digital. Sus principales preocupaciones eran que la paranoia en torno a la IA retrasase la utopía digital y/o provocase una militarización de aquella que chocaría con la divisa de Google: *Don't be evil* («No seas malvado»). Elon no dejó de rebatir sus razonamientos y de pedirle a Larry que aclarase detalles de sus argumentos, como por qué tenía tanta confianza en que la vida digital no destruiría todo lo que consideramos importante. En varias ocasiones, Larry acusó a Elon de ser «especista» por tratar a determinadas formas de vida como inferiores solo porque estaban basadas en el silicio en lugar de en el carbono. Volveremos sobre estas interesantes cuestiones y argumentos en detalle a partir del capítulo 4.

Aunque Larry parecía estar en minoría esa cálida noche estival junto a la piscina, el utopismo digital que con tanta elocuencia defendía cuenta con muchos y muy prominentes defensores. El especialista en robótica y futurólogo Hans Moravec inspiró a toda una generación de utópicos digitales con su clásico de 1988 titulado *El hombre mecánico*, una tradición que continuó y refinó el inventor Ray Kurzweil. Richard Sutton, uno de los pioneros del subcampo de la IA conocido como aprendizaje por refuerzo, hizo una apasionada defensa del utopismo digital en nuestra conferencia en Puerto Rico de la que hablaré enseguida.

Tecnoescépticos

Hay otro grupo prominente de pensadores a los que tampoco preocupa la IA, pero por una razón completamente distinta: creen que construir una IAG

sobrehumana es tan difícil que aún tardará cientos de años en suceder, y por lo tanto consideran ridículo preocuparse por ello ahora. Es lo que denomino la postura *tecnoescéptica*, que Andrew Ng expone de forma elocuente: «Temer la rebelión de los robots asesinos es como preocuparse por la superpoblación en Marte». Andrew era el científico jefe en Baidu, el Google chino, y recientemente repitió este argumento cuando hablé con él en una conferencia celebrada en Boston. También me explicó que creía que preocuparse por los riesgos de la IA era una distracción potencialmente peligrosa que podría demorar el progreso de la misma. Otros tecnoescépticos, como Rodney Brooks (el exprofesor en el MIT que está detrás de la creación de la aspiradora robótica Roomba y el robot industrial Baxter), han expresado opiniones similares. Me parece interesante ver cómo, aunque los utópicos digitales y los tecnoescépticos coinciden en que no deberíamos preocuparnos por la IA, no están de acuerdo en prácticamente nada más. La mayoría de los utópicos piensan que la IAG de nivel humano podría aparecer en un plazo de entre veinte y cien años, algo que los tecnoescépticos desestiman como una ensoñación sin fundamento, y a menudo se mofan de la singularidad que los utópicos profetizan como «el rapto de los *geeks*». Cuando conocí a Rodney Brooks en una fiesta de cumpleaños en diciembre de 2014, me dijo que estaba seguro al cien por cien de que no sucedería en el tiempo que me quedaba de vida. «¿Estás seguro de que no quieres decir 99 %?, le pregunté en un correo electrónico poco tiempo después, al que me respondió así: «Nada de un pusilánime 99 %. Cien por cien. Simplemente no sucederá».

El movimiento en pro de una IA benéfica

Cuando conocí a Stuart Russell en una cafetería parisina en junio de 2014, me pareció el caballero británico por antonomasia. Elocuente, mesurado y de voz suave, pero con un brillo aventurero en los ojos, era la encarnación moderna de Phileas Fogg, mi héroe de infancia, protagonista de *La vuelta al mundo en ochenta días*, la novela clásica que Jules Verne publicó en 1873. Aunque era uno de los más famosos investigadores en IA vivos, y había coescrito el libro de texto de referencia sobre el tema, su modestia y amabilidad enseguida consiguieron que me sintiera cómodo. Me explicó cómo los avances en IA lo habían convencido de que existía una posibilidad

real de que la IAG de nivel humano surgiese a lo largo de este siglo, pero, aunque él era optimista, no estaba garantizado que las consecuencias de ello fuesen positivas. Antes había que responder a algunas preguntas fundamentales, tan difíciles que deberíamos empezar ya a investigar sobre ellas, para tener las respuestas preparadas cuando las necesitásemos.

Hoy en día, las opiniones de Stuart son predominantes, y muchos grupos de investigación de todo el mundo llevan a cabo el tipo de investigación en IA segura que él defiende. Pero no siempre fue así. Un artículo en el *Washington Post* se refirió a 2015 como el año en que la investigación en IA segura se hizo popular. Hasta entonces, cuando alguien hablaba de los riesgos de la IA, los investigadores de las corrientes dominantes en IA lo malinterpretaban y lo despachaban como alarmismo ludita dirigido a impedir el progreso de la IA. Como veremos en el capítulo 5, el pionero de la computación Alan Turing y el matemático Irving J. Good, que colaboró con Turing para descifrar los códigos alemanes durante la Segunda Guerra Mundial, fueron los primeros en expresar, hace más de medio siglo, una inquietud en consonancia con la de Stuart. En la última década, la investigación en estos temas la llevaron a cabo principalmente un puñado de pensadores independientes que no eran investigadores profesionales en IA, como Eliezer Yudkowsky, Michael Vassar y Nick Bostrom. Sus trabajos tuvieron poco eco entre la mayoría de los principales investigadores en IA, que solían centrarse en su trabajo diario de cómo hacer que los sistemas de IA fuesen más inteligentes, en lugar de plantearse las consecuencias a largo plazo de sus avances. Muchos de los investigadores en IA que yo conocía y que sabía que albergaban alguna preocupación no osaban plantearla públicamente por temor a que se los viese como tecnófobos alarmistas.

Me parecía que había que cambiar esa situación polarizada, para que la comunidad de la IA al completo pudiese participar e influir en la conversación sobre cómo construir una IA benéfica. Afortunadamente, no era yo el único que lo creía. En la primavera de 2014, fundé una organización sin ánimo de lucro llamada Future of Life Institute (FLI; <http://futureoflife.org>) junto con mi mujer, Meia, mi amigo el físico Anthony Aguirre, una estudiante de doctorado en Harvard, Viktoriya Krakovna, y el fundador de Skype, Jaan Tallinn. Nuestro objetivo era simple: contribuir a garantizar que la vida tuviera futuro y que este fuera tan fabuloso como fuese posible. En particular, creíamos que la tecnología permitía a la

vida prosperar como nunca antes o bien autodestruirse, y nosotros preferíamos que sucediese lo primero.

Nuestra primera reunión consistió en una sesión de tormenta de ideas en nuestra casa el 15 de marzo de 2014, con alrededor de treinta alumnos, profesores y otros pensadores de la zona de Boston. Hubo un amplio consenso en que, aunque teníamos que prestar atención a la biotecnología, las armas nucleares y el cambio climático, nuestro primer objetivo debería ser conseguir que se generalizase la investigación sobre IA segura. Mi colega en el departamento de Física del MIT Frank Wilczek, que había recibido el Premio Nobel por contribuir a entender cómo se comportan los quarks, sugirió que empezásemos escribiendo un artículo de opinión en el periódico para llamar la atención sobre el asunto y hacer que fuese más difícil ignorarlo. Hablé con Stuart Russell (a quien todavía no conocía en persona) y con mi colega físico Stephen Hawking, y ambos accedieron a figurar como coautores junto a Frank y a mí. Muchas revisiones más tarde, nuestro artículo fue rechazado por el *New York Times* y muchos otros periódicos estadounidenses, por lo que lo publicamos en mi blog en el *Huffington Post.* Para mi gran alegría, la propia Arianna Huffington me escribió diciendo: «¡Encantada de tenerlo! ¡Lo publicaremos en el lugar más destacado!», y esta ubicación en la cabecera de la portada desencadenó una oleada de cobertura mediática de la IA segura que se prolongó durante el resto de ese año, con intervenciones de Elon Musk, Bill Gates y otros líderes tecnológicos. *Superinteligencia,* el libro de Nick Bostrom, se publicó ese otoño y alimentó aún más el creciente debate público.

El siguiente objetivo de la campaña del FLI en pro de la IA benéfica consistió en reunir a los principales investigadores en IA en una conferencia donde se pudieran aclarar los malentendidos, alcanzar algún tipo de consenso y hacer planes constructivos. Sabíamos que sería difícil persuadir a una concurrencia tan ilustre para que asistiese a una conferencia organizada por intrusos en su campo a quienes no conocían, más aún habida cuenta de lo controvertido del tema, por lo que lo intentamos de todas las maneras posibles: prohibimos la asistencia de la prensa, decidimos organizarla en la playa (en Puerto Rico) en enero, conseguimos que la asistencia fuese gratuita (gracias a la generosidad de Jaan Tallinn) y le pusimos el título menos alarmista que se nos ocurrió: «El futuro de la IA: oportunidades y retos». Pero lo más importante fue que aunamos esfuerzos con Stuart Russell.

Gracias a él, logramos que se sumaran al comité organizador un grupo de líderes en IA tanto procedentes del ámbito académico como de la industria, entre los que estaba Demis Hassabis, de DeepMind (empresa propiedad de Google), que posteriormente demostraría que la IA podía vencer a los humanos incluso al juego del go. Cuanto más conocía a Demis, más cuenta me daba de que este ambicionaba no solo hacer que la IA fuese potente, sino también benéfica.

El resultado fue un extraordinario encuentro de cerebros (figura 1.3). A los investigadores en IA se unieron economistas, expertos en Derecho, líderes de la industria tecnológica (incluido Elon Musk), así como otros pensadores (como Vernor Vinge, que acuñó el término «singularidad», sobre el que gira el capítulo 4). El resultado superó incluso nuestras expectativas más optimistas. Puede que fuese la combinación del sol y del vino, o quizá simplemente que era el momento adecuado: a pesar de lo controvertido que resultaba el tema, se alcanzó un consenso notable, que plasmamos en una carta abierta[2] que acabó siendo firmada por más de ocho mil personas, incluida la plana mayor de la IA. El contenido esencial de la carta era que había que redefinir el objetivo de la IA, que no debía ser el de crear una IA sin dirección, sino una benéfica. La carta también mencionaba una lista detallada de líneas de investigación que los participantes en la conferencia coincidían en que promoverían este objetivo. El movimiento en pro de una IA benéfica empezaba a ganar popularidad. Más adelante en el libro daremos cuenta de cuál ha sido su desarrollo desde entonces.



FIGURA 1.3. La conferencia de enero de 2015 en Puerto Rico reunió a un extraordinario grupo de investigadores en IA y otros campos relacionados. Fila posterior, de izquierda a derecha: Tom Mitchell, Seán Ó hÉigeartaigh, Huw Price, Shamil Chandaria, Jaan Tallinn, Stuart Russell, Bill Hibbard, Blaise Agüera y Arcas, Anders Sandberg, Daniel Dewey, Stuart Armstrong, Luke Muehlhauser, Tom Dietterich, Michael Osborne, James Manyika, Ajay Agrawal, Richard Mallah, Nancy Chang, Matthew Putman. Otra gente de pie, de izquierda a derecha: Marilyn Thompson, Rich Sutton, Alex Wissner-Gross, Sam Teller, Toby Ord, Joscha Bach, Katja Grace, Adrian Weller, Heather Roff-Perkins, Dileep George, Shane Legg, Demis Hassabis, Wendell Wallach, Charina Choi, Ilya Sutskever, Kent Walker, Cecilia Tilli, Nick Bostrom, Erik Brynjolfsson, Steve Crossan, Mustafa Suleyman, Scott Phoenix, Neil Jacobstein, Murray Shanahan, Robin Hanson, Francesca Rossi, Nate Soares, Elon Musk, Andrew McAfee, Bart Selman, Michele Reilly, Aaron VanDevender, Max Tegmark, Margaret Boden, Joshua Greene, Paul Christiano, Eliezer Yudkowsky, David Parkes, Laurent Orseau, JB Straubel, James Moor, Sean Legassick, Mason Hartman, Howie Lempel, David Vladeck, Jacob Steinhardt, Michael Vassar, Ryan Calo, Susan Young, Owain Evans, Riva-Melissa Tez, János Krámar, Geoff Anders, Vernor Vinge, Anthony Aguirre. Sentados: Sam Harris, Tomaso Poggio, Marin Soljačić, Viktoriya Krakovna, Meia Chita-Tegmark. Tras la cámara: Anthony Aguirre (que aparece en la foto gracias a la habilidad con Photoshop de la inteligencia de nivel humano que está a su lado).

Otra lección importante de la conferencia fue esta: las cuestiones que suscita el éxito de la IA no son solo intelectualmente fascinantes, también son fundamentales desde un punto de vista moral, porque nuestras decisiones pueden afectar en potencia a todo el futuro de la vida. La transcendencia moral de las decisiones que la humanidad tomó en el pasado fue en ocasiones grande, pero siempre limitada: nos hemos recuperado de las peores plagas, e incluso los mayores imperios han acabado por derrumbarse. Las generaciones pasadas sabían que, tan seguro como que el Sol saldría al día siguiente, los

humanos también seguirían ahí, combatiendo eternas lacras como la pobreza, la enfermedad y la guerra. Pero algunos de los intervinientes en la conferencia de Puerto Rico argumentaron que esta vez podía ser diferente: por primera vez, decían, podríamos construir una tecnología lo suficientemente potente para acabar para siempre con esas lacras… o con la propia humanidad. Podríamos crear sociedades que prosperasen como ninguna lo había hecho antes, en la Tierra y quizá fuera de ella, o un Estado de vigilancia global kafkiano tan poderoso que nunca podría ser derrocado.



FIGURA 1.4. Aunque los medios suelen presentar a Elon Musk como si estuviese enfrentado con la comunidad de la IA, en realidad existe un amplio consenso en torno a la necesidad de la investigación en IA segura. En la foto, tomada el 4 de enero de 2015, Tom Dietterich, presidente de la Asociación para el Avance de la Inteligencia Artificial, comparte la ilusión de Elon por el nuevo programa de investigación en IA segura que este último se acababa de comprometer a financiar unos minutos antes. Tras ellos, aparecen las fundadoras del FLI Meia Chita-Tegmark y Viktoriya Krakovna.

MALENTENDIDOS

Cuando salí de Puerto Rico, lo hice convencido de que la conversación que habíamos tenido allí sobre el futuro de la IA debía continuar, porque es la conversación más importante de nuestro tiempo.(3) Es una conversación que trata sobre el futuro colectivo de todos nosotros, por lo que en ella no deberían participar únicamente los investigadores en IA. Por eso escribí este libro. Lo hice con la esperanza de que usted, mi estimado lector, se una a la conversación. ¿Qué tipo de futuro quiere? ¿Deberíamos desarrollar armas letales autónomas? ¿Cómo querría usted que evolucionase la automatización del trabajo? ¿Qué consejos les daría a los jóvenes de hoy sobre su futuro profesional? ¿Prefiere que haya nuevos trabajos que sustituyan a los antiguos, o que vivamos en una sociedad sin trabajo donde todo el mundo disfrute de una vida de ocio y sean las máquinas las que produzcan la riqueza? Más adelante, ¿querría usted que creásemos la vida 3.0 y la difundiésemos por el cosmos? ¿Controlaremos a las máquinas inteligentes o serán ellas las que nos controlen a nosotros? Esas máquinas inteligentes ¿nos sustituirán, coexistirán con nosotros o se fusionarán con nosotros? ¿Qué significará ser humano en la era de la inteligencia artificial? ¿Qué querría que significase, y cómo podemos hacer que ese sea nuestro futuro?

El objetivo de este libro es ayudarle a participar en esta conversación. Como ya he comentado, existen fascinantes controversias entre los mayores expertos mundiales. Pero también he visto muchos ejemplos de aburridas pseudocontroversias en las que las personas se malinterpretan unas a las otras y mantienen diálogos de sordos. Para ayudar a centrarnos en las polémicas interesantes y en las cuestiones abiertas, y no en los malentendidos, comencemos por aclarar algunas de las ideas erróneas más difundidas.

Existen numerosas definiciones contrapuestas de uso habitual para términos como «vida», «inteligencia» o «consciencia», y muchas interpretaciones erróneas se deben a que la gente no sabe que está usando palabras con dos sentidos distintos. Para asegurarnos de que usted y yo no caemos en esta trampa, en la tabla 1.1 he incluido una ficha de referencia donde se explica cómo uso las expresiones clave de este libro. Algunas de estas definiciones solo se introducirán y se explicarán debidamente en capítulos posteriores. Por favor, tenga en cuenta que no quiero dar a entender que mis definiciones sean mejores que las de cualquier otro; solo trato de evitar la confusión al precisar lo que quiero decir. Verá que por lo general

opto por definiciones amplias que evitan el sesgo antropocéntrico y que pueden aplicarse tanto a máquinas como a humanos. Por favor, lea la ficha ahora, y vuelva a ella más adelante si le desconcierta mi uso de algunas de las palabras que contiene (en particular en los capítulos 4 a 8).

FICHA DE TERMINOLOGÍA	
Vida	Proceso capaz de preservar su complejidad y de replicarse
Vida 1.0	Vida cuyo hardware y software son resultado de la evolución (fase biológica)
Vida 2.0	Vida cuyo hardware es resultado de la evolución, pero que diseña buena parte su software (fase cultural)
Vida 3.0	Vida que diseña su hardware y su software (fase tecnológica)
Inteligencia	Capacidad de alcanzar objetivos complejos
Inteligencia artificial (IA)	Inteligencia no biológica
Inteligencia estrecha (o débil)	Capacidad de alcanzar un conjunto limitado de objetivos; por ejemplo, jugar al ajedrez o conducir un coche
Inteligencia general	Capacidad de alcanzar prácticamente cualquier objetivo, incluido el aprendizaje
Inteligencia universal	Capacidad de adquirir inteligencia general a partir del acceso a datos y recursos
Inteligencia artificial general (IAG)	Capacidad para realizar cualquier tarea cognitiva al menos tan bien como los humanos
IA de nivel humano	IAG
IA fuerte	IAG
Superinteligencia	Inteligencia general de nivel muy superior al humano
Civilización	Grupo de formas de vida inteligente que interactúan entre sí
Consciencia	Experiencia subjetiva
Qualia	Casos individuales de experiencia subjetiva
Ética	Principios que rigen cómo deberíamos comportarnos
Teleología	Explicación de las cosas en función de sus objetivos o propósitos, y no de sus causas
Comportamiento orientado a objetivos	Comportamiento que se explica más fácilmente a través de su efecto que de su causa
Tener un objetivo	Mostrar un comportamiento orientado a objetivos
Tener un propósito	Servir a los objetivos propios o de otra entidad
IA amigable	Superinteligencia cuyos objetivos coinciden con los nuestros

Cíborg	Híbrido hombre-máquina
Explosión de inteligencia	Automejora recursiva rápidamente conducente a superinteligencia
Singularidad	Explosión de inteligencia
Universo	La región de espacio desde la que la luz ha tenido tiempo para llegar hasta nosotros durante los 13.800 millones de años transcurridos desde el Big Bang

TABLA 1.1. Muchos malentendidos en torno a la IA se deben a que la gente usa las palabras que se recogen en esta tabla para referirse a cosas distintas. Aquí expongo el significado que tienen para mí en este libro. (Algunas de estas definiciones solo se introducirán y explicarán debidamente en capítulos posteriores.)

Además de la confusión en torno a la terminología, también he visto cómo muchas conversaciones sobre IA se frustran por culpa de simples malentendidos. Aclaremos algunos de las más comunes.

Mitos sobre la cronología

El primero tiene que ver con la cronología de la figura 1.2: ¿cuánto tiempo transcurrirá hasta que las máquinas superen ampliamente la IAG de nivel humano? Aquí, un error habitual consiste en pensar que conocemos la respuesta.

Un mito muy extendido es pensar que conseguiremos IAG sobrehumana en este siglo. De hecho, la historia está repleta de ejemplos de exagerado optimismo tecnológico. ¿Dónde están esas centrales de fusión nuclear y coches voladores que nos prometieron que tendríamos a estas alturas? La IA también ha padecido exceso de bombo una y otra vez en el pasado, incluso por parte de algunos de los fundadores del campo: John McCarthy (que acuñó la expresión «inteligencia artificial»), Marvin Minsky, Nathaniel Rochester y Claude Shannon hicieron una estimación excesivamente optimista de lo que podría logarse en dos meses con ordenadores de la edad de piedra: «Proponemos que durante el verano de 1956 se lleve a cabo un estudio sobre inteligencia artificial de dos meses de duración y que ocupe a diez personas en Dartmouth College […]. Se intentará encontrar la manera de hacer que las máquinas usen el lenguaje, formen abstracciones y conceptos, resuelvan tipos de problemas hasta ahora reservados a los humanos y se

mejoren a sí mismas. Pensamos que se pueden hacer avances significativos en uno o más de estos problemas si un grupo cuidadosamente seleccionado de científicos trabajan conjuntamente en ello durante un verano».

Por otra parte, un mito popular contrario es creer que no tendremos IAG sobrehumana este siglo. Los investigadores han establecido un amplio abanico de estimaciones del tiempo que tardaremos en tener IAG sobrehumana, pero es cierto que no se puede asegurar que la probabilidad para este siglo sea cero, dado el pobre historial de predicciones tecnoescépticas como esta. Por ejemplo, Ernest Rutherford, posiblemente el más destacado físico nuclear de su época, dijo en 1933 —menos de veinticuatro horas antes de que Leo Szilard inventase la reacción nuclear en cadena— que la energía nuclear eran «pamplinas», y en 1956 el Astrónomo Real Richard Woolley se refirió a los rumores sobre viajes al espacio como «una soberana sandez». La forma más extrema de este mito es que la IAG sobrehumana nunca llegará porque es físicamente imposible, a pesar de que los físicos saben que un cerebro está compuesto por quarks y electrones dispuestos de tal manera que funcionan como un potente ordenador, y que no existe ninguna ley física que nos impida construir masas de quarks todavía más inteligentes.

Se han hecho toda una serie de sondeos entre los investigadores en IA en los que se les preguntaba dentro de cuántos años creían que tendríamos IAG de nivel humano con una probabilidad de al menos el 50 %, y todos ellos llegaron a la misma conclusión: los mayores expertos mundiales en IA no se ponen de acuerdo, así que simplemente no lo sabemos. Por ejemplo, en uno de estos sondeos realizado entre los investigadores en IA presentes en la conferencia de Puerto Rico, la respuesta promedio (la mediana) era el año 2055, pero algunos de los investigadores preveían que tardaría en llegar varios siglos o incluso más.

Hay otro mito relacionado según el cual las personas a las que les preocupa la IA creen que llegará en tan solo unos pocos años. De hecho, la mayoría de la gente que ha expresado públicamente su preocupación por la IAG sobrehumana cree que aún tardará décadas en llegar. Pero argumentan que, mientras no estemos seguros al cien por cien de que no se producirá durante este siglo, es sensato empezar ahora a investigar sobre seguridad para

prepararnos para tal eventualidad. Como veremos a lo largo del libro, muchos de los problemas de seguridad son tan complicados que podríamos tardar décadas en resolverlos, por lo que es prudente empezar a investigar ya sobre ellos, en lugar de hacerlo la noche antes de que unos programadores hasta arriba de Red Bull decidan poner en marcha la IAG de nivel humano.

Mitos sobre la controversia

Otro error común consiste en pensar que las únicas personas que albergan alguna inquietud acerca de la IA y abogan por la investigación en IA segura son luditas que no saben mucho sobre IA. Cuando Stuart Russell lo mencionó en su charla en Puerto Rico, el público se echó a reír. Otro error relacionado es que apoyar la investigación en IA segura es algo enormemente controvertido. De hecho, para apoyar una modesta inversión en investigación en IA segura, no es necesario estar convencidos de que los riesgos son altos, basta con que estos no sean despreciables, de la misma manera en que una modesta inversión en un seguro de hogar está justificada por una probabilidad no despreciable de que el hogar sea pasto de las llamas.

Mi análisis personal es que los medios de comunicación han hecho que el debate en torno a la IA segura parezca más controvertido de lo que es. A fin de cuentas, el miedo vende, y los artículos que usan citas sacadas de contexto para anunciar un cataclismo inminente pueden generar más clics que otros más matizados y equilibrados. En consecuencia, es probable que dos personas que solo conocen las opiniones de la otra a través de las citas aparecidas en la prensa crean que discrepan más de lo que en realidad lo hacen. Por ejemplo, un tecnoescéptico que solo conozca la postura de Bill Gates a través de un tabloide británico podría pensar erróneamente que este cree que la superinteligencia es un fenómeno inminente. De forma análoga, alguien en el movimiento en pro de una IA benéfica que no conozca sobre la postura de Andrew Ng otra cosa que la cita antes mencionada sobre la superpoblación en Marte podría creer que no le preocupa la seguridad en IA. De hecho, sé que no es así: la clave está en que, dado que sus estimaciones prevén plazos más largos, Andrew tiende de manera natural a priorizar los problemas de la IA a corto plazo frente a los de más largo plazo.

Mitos sobre cuáles son los riesgos

Me exasperó leer este titular en el *Daily Mail*: «Stephen Hawking advierte de que la irrupción de los robots puede ser desastrosa para la humanidad».[3] He perdido la cuenta de cuántos artículos como este he visto. Suelen ir acompañados de la imagen de un robot de aspecto perverso y con un arma, y sugieren que debería inquietarnos que los robots se rebelen y nos maten porque se han vuelto conscientes y/o malvados. En un tono menos serio, hay que decir que ese tipo de artículos resultan en realidad bastante impresionantes, porque resumen de forma sucinta el escenario que a mis colegas que trabajan en IA no les preocupa. Dicho escenario es la combinación de hasta tres malentendidos distintos, que tienen que ver con la *consciencia*, la *maldad* y los *robots*.

Cuando conducimos por una carretera, tenemos una experiencia subjetiva de los colores, los sonidos, etcétera. Pero ¿tiene un coche autónomo una experiencia subjetiva? ¿Se siente algo al ser un coche autónomo, o es como ser un zombi inconsciente sin ninguna experiencia subjetiva? Aunque este misterio de la consciencia es interesante por sí mismo, y a él dedicaremos el capítulo 8, es irrelevante para el riesgo que pueda suponer la IA. Si nos atropella un coche autónomo, da igual que se sienta subjetivamente consciente o no. En ese mismo sentido, lo que nos afectará a los humanos es lo que haga la IA superinteligente, no lo que sienta desde un punto de vista subjetivo.

El temor a que las máquinas se vuelvan malévolas es otra trampa. Lo realmente preocupante no es la malevolencia, sino la competencia. Una IA superinteligente es, por definición, muy buena a la hora de alcanzar sus objetivos, sean los que sean, por lo que necesitamos asegurarnos de que coinciden con los nuestros. Es probable que usted no odie a las hormigas y vaya por ahí pisoteándolas por mera maldad, pero si está a cargo de un proyecto de energía verde hidroeléctrica y resulta que hay un hormiguero en la zona que quedará bajo el agua, mala suerte para las hormigas. El movimiento en pro de una IA benéfica busca evitar que la humanidad acabe ocupando el lugar de esas hormigas.

El malentendido sobre la consciencia está relacionado con el mito de que las máquinas no pueden tener objetivos. Es evidente que sí pueden tenerlos

en el sentido limitado de que pueden exhibir un comportamiento orientado hacia la consecución de objetivos: la forma más económica de explicar el comportamiento de un misil guiado por el calor consiste en decir que tiene por objetivo dar en el blanco. Si nos sentimos amenazados por una máquina cuyos objetivos no coinciden con los nuestros, es porque son estos objetivos, en el sentido más limitado, lo que nos preocupa, no el hecho de que la máquina sea consciente y experimente la sensación de tener un propósito. Si ese misil guiado por el calor nos persiguiese, probablemente no exclamaríamos «¡No tengo miedo, porque las máquinas no pueden tener objetivos!».

Simpatizo con Rodney Brooks y otros pioneros de la robótica que se sienten injustamente demonizados por los tabloides alarmistas, porque algunos periodistas parece que tienen una fijación obsesiva con los robots e ilustran muchos de sus artículos con monstruos metálicos con brillantes ojos rojos y aspecto malvado. De hecho, la principal preocupación del movimiento en pro de una IA benéfica no son los robots sino la inteligencia en sí: en particular, la inteligencia cuyos objetivos no coincidan con los nuestros. Para causarnos problemas, esta inteligencia mal orientada no necesita ningún cuerpo robótico, sino tan solo una conexión a internet. En el capítulo 4 veremos cómo esto puede permitir ganarles la partida a los mercados financieros, ser más ingeniosos que los investigadores humanos, adelantarse a las maniobras de los líderes políticos humanos, y desarrollar armas que nosotros ni siquiera podamos entender. Incluso si construir robots fuese físicamente imposible, una IA superinteligente y superrica podría pagar o manipular a un sinfín de humanos para que cumplieran sus órdenes como sucede en *Neuromancer*, la novela de ciencia ficción de William Gibson.

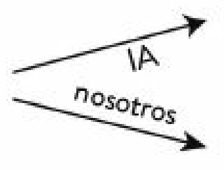
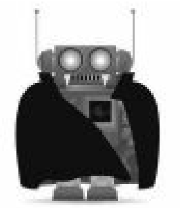
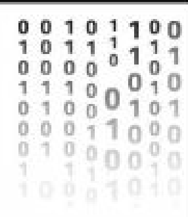
MITO La superinteligencia antes de 2100 es inevitable MITO La superinteligencia antes de 2100 es imposible	VERDAD Puede que llegue en décadas, siglos o nunca: los expertos en IA no se ponen de acuerdo, y nosotros simplemente no lo sabemos 
MITO Solo a los luditas les preocupa la IA 	VERDAD A muchos de los investigadores más importantes en IA les preocupa 
PREOCUPACIÓN IMAGINARIA Que la IA se vuelva malvada Preocupación imaginaria Que la IA adquiera consciencia 	PREOCUPACIÓN REAL Que la IA llegue a ser competente y tenga objetivos no alineados con los nuestros 
MITO Los robots son la principal fuente de preocupación 	VERDAD La inteligencia mal orientada es la principal inquietud: no necesita ningún cuerpo, solo una conexión a internet 
MITO La IA no puede controlar a los humanos 	VERDAD La inteligencia posibilita el control: nosotros controlamos a los tigres porque somos más inteligentes 
MITO Las máquinas no pueden tener objetivos 	VERDAD Un misil que se guía por el calor tiene un objetivo 
PREOCUPACIÓN IMAGINARIA Quedan solo unos años para que llegue la superinteligencia  	PREOCUPACIÓN REAL Quedan al menos varias décadas para que llegue, pero puede que necesitemos todo ese tiempo para hacer que sea segura 

FIGURA 1.5. Mitos extendidos sobre la IA superinteligente.

El malentendido en relación con los robots tiene que ver con el mito de que las máquinas no pueden controlar a los humanos. La inteligencia hace posible el control: los humanos controlamos a los tigres no porque seamos más fuertes, sino porque somos más inteligentes. Esto significa que, si dejamos de ser los más inteligentes del planeta, es posible que perdamos también el control.

La figura 1.5 resume todos estos malentendidos habituales, para que podamos olvidarnos de ellos de una vez por todas y centrar nuestras discusiones con amigos y colegas en las muchas controversias que existen realmente (como veremos, no escasean).

EL CAMINO POR RECORRER

En el resto del libro, usted y yo exploraremos juntos el futuro de la vida con IA. Recorreremos este tema rico y poliédrico de forma organizada, empezando por explorar conceptual y cronológicamente la historia completa de la vida, después pasaremos a reflexionar sobre qué entendemos por objetivos y significado, y terminaremos viendo qué acciones tomar para crear el futuro que queremos.

En el capítulo 2, exploraremos los fundamentos de la inteligencia y cómo la materia en apariencia tonta puede reorganizarse para recordar, computar y aprender. A medida que nos adentramos en el futuro, nuestra historia se ramifica en muchos escenarios definidos por las respuestas a determinadas preguntas clave. La figura 1.6 resume las cuestiones clave con las que nos encontraremos a medida que avancemos en el tiempo hacia una IA potencialmente más y más avanzada.

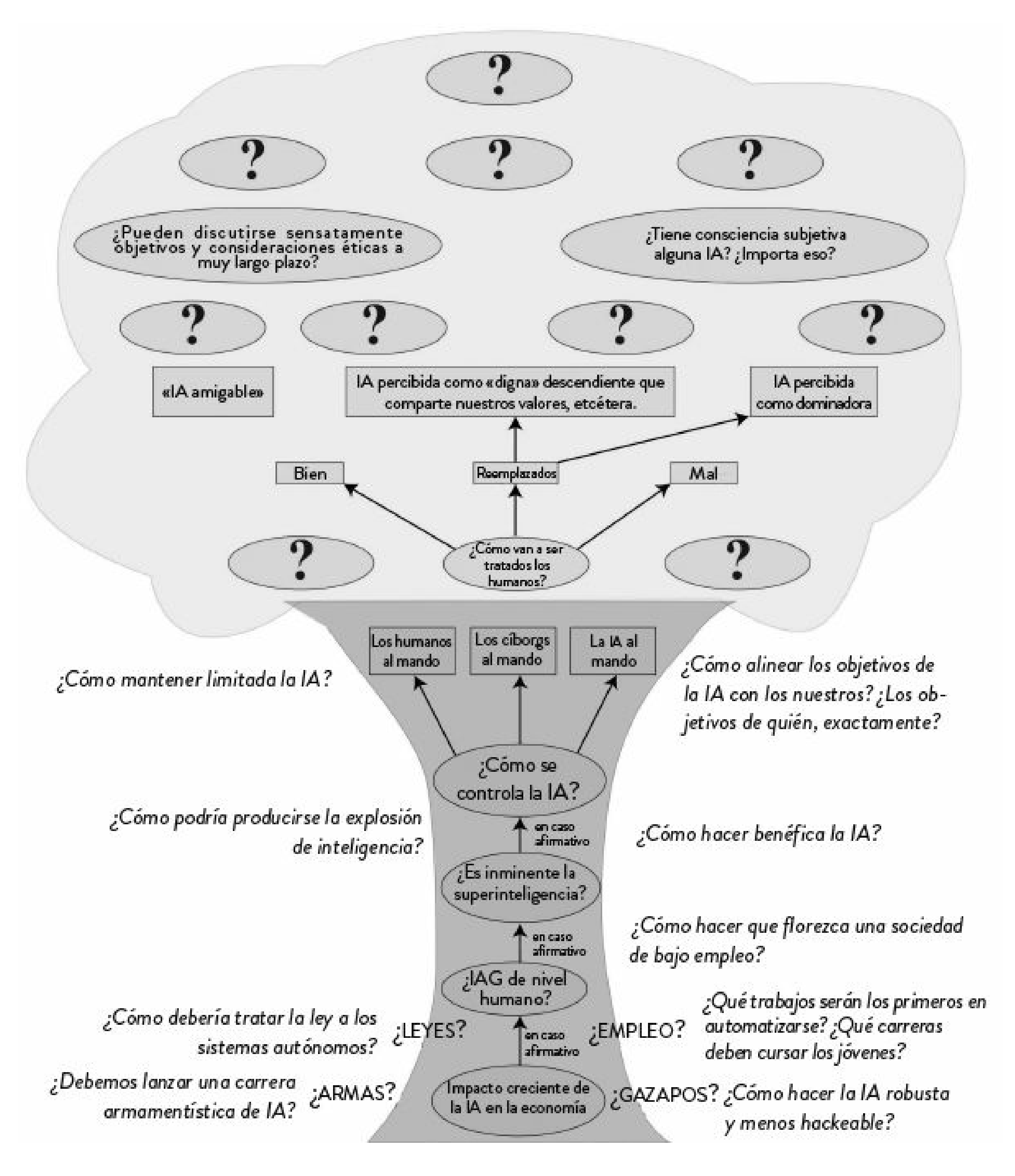


FIGURA 1.6. Qué cuestiones en torno a la IA son interesantes depende de hasta dónde avance y por cuál de las ramas se desarrollará nuestro futuro.

Ahora mismo, hemos de hacer frente al dilema de si iniciar o no una carrera armamentística en IA, y a cuestiones relativas a cómo conseguir que los sistemas de IA del mañana sean robustos y perfectos. Si el impacto

económico de la IA continúa creciendo, también tendremos que decidir cómo modernizar nuestras leyes y cómo orientar profesionalmente a nuestros jóvenes para que eviten los trabajos que en poco tiempo serán automatizados. Exploramos estas cuestiones relativas al corto plazo en el capítulo 3.

Si el progreso de la IA continúa hasta alcanzar niveles humanos, también tendremos que plantearnos cómo asegurarnos de que es benéfica, y si podemos o debemos crear una sociedad del ocio que prospere sin trabajos. Esto también suscita la cuestión de si una explosión de inteligencia o una progresión lenta pero constante puede impulsar la IAG mucho más allá de los niveles humanos. En el capítulo 4 exploramos una amplia variedad de escenarios de este estilo e investigamos el espectro de posibilidades que podrían darse en el capítulo 5, que van desde lo presumiblemente distópico hasta lo posiblemente utópico. ¿Quién está al mando: los humanos, la IA o los cíborgs? ¿Los humanos reciben buen trato, o no? ¿Somos sustituidos y, de ser así, percibimos a nuestros sustitutos como conquistadores o como dignos descendientes? Tengo curiosidad por saber cuál de los escenarios que se plantean en el capítulo 5 prefiere usted. He montado un sitio web, <http://AgeOfAi.org>, donde puede expresar su opinión y participar en la conversación.

Por último, en el capítulo 6 nos adentramos miles de millones de años en el futuro, cuando los límites últimos de la vida en el cosmos no dependen de la inteligencia, sino de las leyes físicas, lo que nos permitirá, paradójicamente, extraer conclusiones más firmes.

Terminada nuestra exploración de la historia de la inteligencia, dedicaremos el resto del libro a considerar qué futuro es preferible y cómo hacer que suceda. Para relacionar los fríos datos con las cuestiones relativas al propósito y al significado, analizamos la base física de los objetivos en el capítulo 7, y la de la consciencia en el capítulo 8. En el epílogo exploramos qué podemos hacer desde ahora mismo para contribuir a crear el futuro que queremos.

		TÍTULO ABREVIADO DEL CAPÍTULO	TEMA	GRADO DE ESPECULACIÓN
		Prólogo. La historia del equipo Omega	Cuestiones que considerar	Sumamente especulativo
La historia de la inteligencia	1	La conversación	Ideas clave, terminología	No muy especulativo
	2	La materia se vuelve inteligente	Fundamentos de la inteligencia	
	3	IA, economía, armas y leyes	El futuro próximo	
	4	¿Explosión de inteligencia?	Escenarios de superinteligencia	Sumamente especulativo
	5	Después de la explosión	Los siguientes 10.000 años	
	6	Nuestra herencia cósmica	Los siguientes miles de millones de años	
La historia del significado	7	Objetivos	Historia del comportamiento orientado a objetivos	No muy especulativo
	8	Consciencia	Consciencia natural y artificial	Especulativo
		Epílogo. La historia del equipo del FLI	¿Qué deberíamos hacer?	No muy especulativo

FIGURA 1.7. Estructura del libro.

Si es usted uno de esos lectores a los que les gusta saltar de un sitio a otro del libro, sepa que la mayoría de los capítulos son relativamente autocontenidos una vez que haya asimilado la terminología y las definiciones de este primer capítulo y del comienzo del siguiente. Si es usted investigador en IA, puede optar por saltarse todo el capítulo 2, salvo las definiciones iniciales de lo que entenderemos por inteligencia. Si no tiene conocimientos sobre IA, los capítulos 2 y 3 le proporcionarán los argumentos para entender por qué los capítulos 4 al 6 no se pueden ignorar sin más como ciencia ficción irrealizable. La figura 1.7 resume dónde se sitúan los distintos capítulos en el espectro que va desde lo fáctico hasta lo especulativo.

Nos espera un viaje fascinante. ¡Adelante!

CONCLUSIONES

- La vida, definida como un proceso capaz de preservar su complejidad y de replicarse, puede desarrollarse a través de estas tres fases: una fase biológica (1.0), en la que tanto su software como su hardware son producto de la evolución, una fase cultural (2.0), en la que puede diseñar su software (mediante el aprendizaje), y una fase tecnológica (3.0), en la cual puede diseñar también su hardware, convirtiéndose en dueña de su propio destino.
- La inteligencia artificial podría permitirnos crear vida 3.0 a lo largo de este siglo, y está teniendo lugar una fascinante conversación en torno al futuro al que deberíamos aspirar y a cómo podría conseguirse. Existen tres campos principales en este debate: los tecnoescépticos, los utópicos digitales y el movimiento en pro de una IA benéfica.
- Los tecnoescépticos creen que construir una IAG sobrehumana es tan difícil que no sucederá en

cientos de años, por lo que es estúpido preocuparse por ello (y por la vida 3.0) ahora.

- Los utópicos digitales consideran que es probable que suceda a lo largo de este siglo y ven con muy buenos ojos la llegada de la vida 3.0, que para ellos es el siguiente paso, natural y deseable, de la evolución cósmica.
- El movimiento en pro de una IA benéfica también considera que es probable que llegue a lo largo de este siglo, pero no da por descontado que el resultado sea positivo, sino que cree que requiere un importante esfuerzo en la forma de investigación sobre IA segura.
- Más allá de este legítimo debate, en el que los expertos mundiales mantienen opiniones discordantes, hay también aburridas pseudocontroversias debidas a malentendidos. Por ejemplo, no merece la pena perder tiempo discutiendo sobre la «vida», la «inteligencia» o la «consciencia» sin asegurarse antes de que nuestro interlocutor usa estas palabras con el mismo significado. Este libro usa las definiciones recogidas en la tabla 1.1.
- También hay que tener cuidado con las ideas erróneas de la figura 1.5: «La superinteligencia antes de 2100 es inevitable/imposible», «Solo a los luditas les preocupa la IA», «Lo preocupante es que la IA se vuelva malévola y/o consciente, y para ello quedan solo unos pocos años», «Los robots son la principal fuente de preocupación», «La IA no puede controlar a los humanos ni tener objetivos».
- En los capítulos 2 a 6, exploraremos la historia de la inteligencia desde sus humildes orígenes, hace miles de millones de años, hasta posibles futuros cósmicos dentro de miles de millones de años. Primero investigaremos los desafíos a corto plazo, como el empleo, las armas con IA y la búsqueda de la IAG de nivel humano, y posteriormente exploraremos el fascinante espectro de futuros posibles con máquinas inteligentes y/o humanos. ¡Me pregunto qué opciones preferirá usted!
- En los capítulos 7 a 9, pasaremos de frías descripciones fácticas a una exploración de los objetivos, de la consciencia y del significado, e investigaremos lo que podemos hacer ahora mismo para contribuir a crear el futuro que deseamos.
- Considero que esta conversación sobre el futuro de la vida con IA es la más importante de nuestro tiempo. Por favor, ¡súmese a ella!

2
LA MATERIA SE VUELVE INTELIGENTE

Transcurrido un tiempo suficiente, el hidrógeno se transforma en personas.

EDWARD ROBERT HARRISON, 1995

Uno de los acontecimientos más espectaculares de los 13.800 millones años transcurridos desde el Big Bang es que la materia tonta e inerte se ha vuelto inteligente. ¿Cómo pudo suceder y cuánto más inteligentes pueden llegar a ser las cosas en el futuro? ¿Qué tiene la ciencia que decir sobre la historia y el porvenir de la inteligencia en el cosmos? Para poder abordar estas cuestiones, dedicaremos este capítulo a explorar los cimientos y componentes fundamentales de la inteligencia. ¿Qué significa decir que un pedazo de materia es inteligente? ¿Qué implica afirmar que un objeto puede recordar, computar y aprender?

¿QUÉ ES LA INTELIGENCIA?

Recientemente, mi mujer y yo tuvimos ocasión de asistir a un simposio sobre inteligencia artificial organizado por la Fundación Nobel y, cuando a un grupo de destacados investigadores sobre IA se les pidió que definieran la inteligencia, discutieron largo y tendido sin alcanzar un consenso. A nosotros nos pareció muy gracioso: no existe acuerdo alguno sobre lo que es la inteligencia, ni siquiera entre inteligentes investigadores sobre inteligencia. Así pues, está claro que no existe una definición «correcta» e indiscutible para este término. Lo que hay son muchas definiciones candidatas, que incluyen la capacidad para la lógica, la comprensión, la planificación, el conocimiento emocional, la autoconciencia, la creatividad, la resolución de problemas y el aprendizaje.

En nuestra exploración del futuro de la inteligencia, queremos adoptar un punto de vista lo más amplio e inclusivo posible, que no se limite a los tipos

de inteligencia que existen hasta ahora. Por este motivo, la definición que propuse en el capítulo anterior, y la manera en que usaré la palabra a lo largo del libro, es muy amplia:

inteligencia = capacidad de alcanzar objetivos complejos

Esta definición es lo suficientemente amplia para incluir todas las otras mencionadas más arriba, ya que la comprensión, la autoconciencia, la resolución de problemas, el aprendizaje y demás son ejemplos de posibles objetivos complejos. Es también lo bastante amplia para englobar la definición del *Oxford Dictionary* —«la capacidad de adquirir y aplicar conocimiento y habilidades»— ya que uno puede tener como objetivo aplicar conocimiento y habilidades.

Puesto que existen numerosos y distintos objetivos, también hay muchos tipos posibles de inteligencia. Según nuestra definición, no tiene sentido por lo tanto cuantificar la inteligencia de humanos, animales no humanos y máquinas mediante un único número, como el cociente intelectual.(4) ¿Qué es más inteligente: un programa informático que solo es capaz de jugar al ajedrez, o uno que solo sabe jugar al go? No hay ninguna respuesta razonable para esta pregunta, ya que hacen bien cosas distintas que no pueden compararse directamente. Podemos, no obstante, afirmar que un tercer programa es más inteligente que los otros dos si es al menos tan bueno como ellos a la hora de lograr todos los objetivos, y mejor para al menos uno de ellos (jugar al ajedrez, por ejemplo).

Tampoco tiene mucho sentido discutir sobre si algo es o no es inteligente en casos límite, ya que esta capacidad es gradual y no por fuerza un rasgo binario. ¿Qué personas son capaces de hablar? ¿Los recién nacidos? No. ¿Los locutores de radio? Sí. Pero ¿y los niños pequeños capaces de pronunciar diez palabras? ¿O quinientas palabras? ¿Dónde pondríamos el límite? He utilizado deliberadamente la ambigua palabra «complejos» en la definición anterior porque no es muy interesante intentar trazar un límite artificial entre la inteligencia y la no inteligencia, y resulta más útil cuantificar el grado de capacidad para lograr diferentes objetivos.

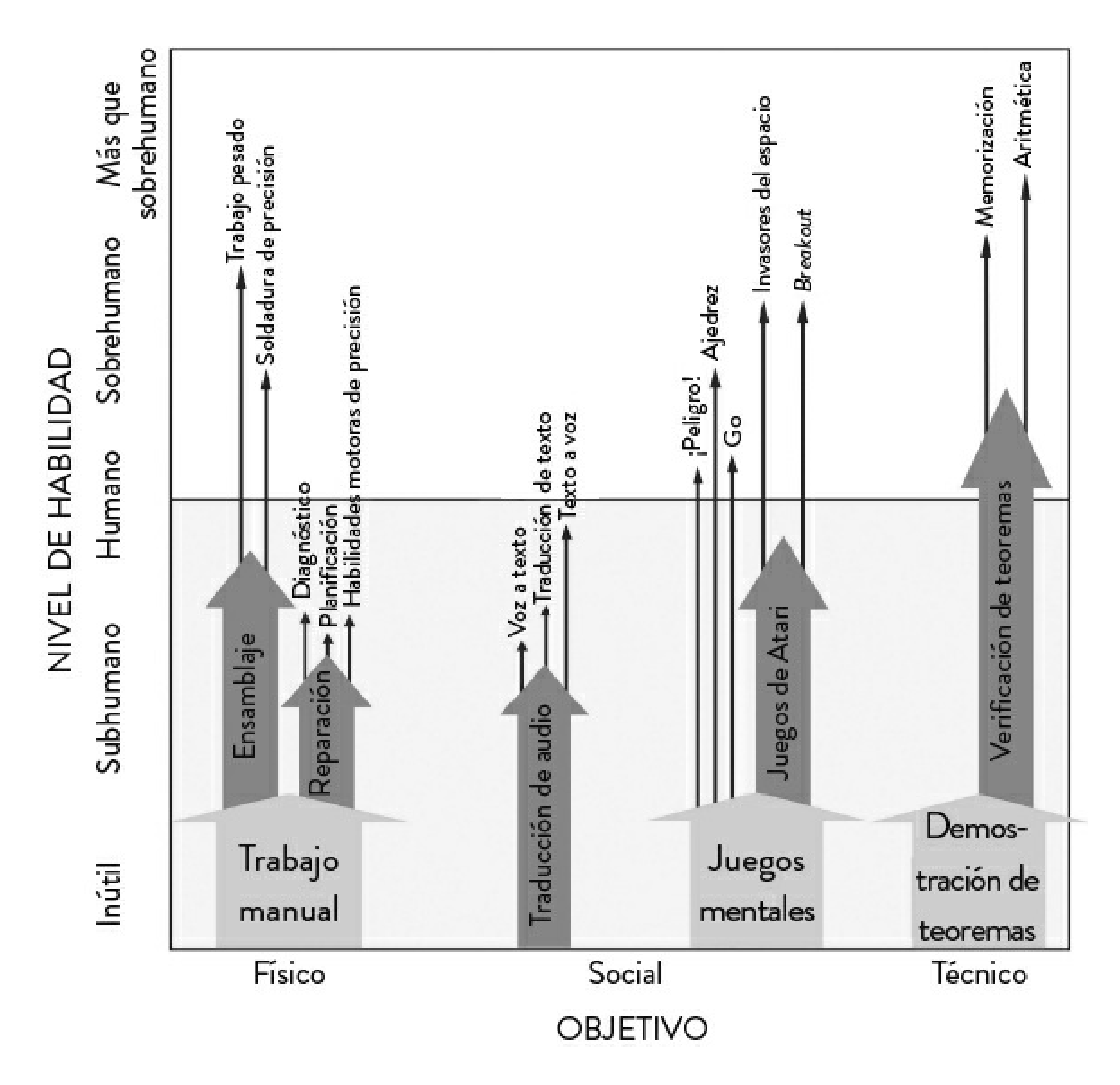


FIGURA 2.1. La inteligencia, definida como la capacidad de lograr objetivos complejos, no puede medirse mediante un único CI, sino solo como un espectro de capacidades en relación con todos los objetivos. Cada flecha indica el grado de habilidad actual de los mejores sistemas de IA para lograr los diferentes objetivos, lo cual pone de manifiesto que la inteligencia artificial actual tiende a ser estrecha: cada sistema es capaz de lograr solo objetivos muy específicos. Por el contrario, la inteligencia humana es extraordinariamente amplia: un niño saludable puede aprender cómo mejorar en casi cualquier cosa.

A la hora de clasificar distintas inteligencias en una taxonomía, otra distinción fundamental es la que existe entre inteligencia *estrecha* y *amplia*. Deep Blue, el ordenador de IBM que juega al ajedrez y que destronó al campeón Garri Kaspárov en 1997, solo era capaz de jugar al ajedrez; a pesar

de estar dotado de un software y de un hardware impresionantes, no podría ganar al tres en raya ni a un niño de cuatro años. El sistema de inteligencia artificial DQN de Google DeepMind es capaz de realizar una variedad de tareas un poco más amplia: puede jugar a decenas de antiguos juegos de ordenador de Atari, como lo haría una persona o incluso mejor. Por el contrario, la inteligencia humana es, hasta la fecha, singularmente amplia, capaz de dominar una impresionante panoplia de habilidades. Un niño saludable, si dispone del suficiente tiempo de formación, puede llegar a desenvolverse no solo en cualquier juego sino también en cualquier idioma, deporte o afición. Si comparamos la inteligencia de los humanos y de las máquinas hoy en día, los humanos ganamos holgadamente en amplitud, mientras que las máquinas nos superan en un conjunto pequeño pero creciente de dominios estrechos, tal y como se ilustra en la figura 2.1. El santo grial de la investigación en inteligencia artificial es construir una «IA general» (más conocida como *inteligencia artificial general*, IAG) de la máxima amplitud: capaz de lograr casi cualquier objetivo, incluido el de aprender. Lo exploraremos en detalle en el capítulo 4. La expresión «IAG» la popularizaron los investigadores Shane Legg, Mark Gubrud y Ben Goertzel para referirse en concreto a una inteligencia artificial general *de nivel humano*: la capacidad de lograr cualquier objetivo al menos tan bien como nosotros.[4] Me ajustaré a su definición, así que, a menos que califique explícitamente el acrónimo (escribiendo «IAG sobrehumana», por ejemplo), usaré «IAG» como abreviatura de «IAG de nivel humano».(5)

Aunque la palabra «inteligencia» tiene por lo general connotaciones positivas, es importante señalar que aquí la estamos usando de una manera completamente neutra: como la capacidad de alcanzar objetivos complejos con independencia de si dichos objetivos se consideran buenos o malos. Así pues, una persona inteligente puede ser muy buena a la hora de ayudar a otras personas, o tener gran capacidad de hacerles daño. Exploraremos esta cuestión de los objetivos en el capítulo 7. Debemos también dejar bien claro cuáles son estos objetivos de los que hablamos. Supongamos que nuestro futuro y flamante asistente personal robótico no tiene ningún objetivo propio pero hará todo lo que le ordenemos, y le pedimos que cocine la cena italiana perfecta. Si se conecta a internet para buscar recetas italianas, para saber cómo llegar al supermercado más próximo o cómo colar la pasta, etcétera, y a continuación logra comprar los ingredientes y preparar una suculenta comida,

es muy probable que lo consideremos inteligente a pesar de que el objetivo original era nuestro. De hecho, el robot asumió nuestro objetivo una vez que hicimos nuestra solicitud, y a continuación lo descompuso en una jerarquía de subobjetivos propios, desde pagar a la cajera hasta rallar el parmesano. En este sentido, el comportamiento inteligente está inextricablemente ligado al logro de objetivos.

Para los humanos es natural clasificar la dificultad de las tareas en función de lo difícil que es para nosotros realizarlas, como en la figura 2.1. Pero esto puede dar una idea engañosa de lo difíciles que son para los ordenadores. Parece mucho más difícil multiplicar 314.159 por 271.828 que reconocer a un amigo en una foto, y sin embargo, los ordenadores nos aplastaron en aritmética mucho antes de que yo naciese, mientras que el reconocimiento de imágenes al nivel humano solo es posible desde hace muy poco tiempo. El hecho de que las tareas sensomotoras de bajo nivel parezcan fáciles a pesar de que requieran una enorme cantidad de recursos se conoce como la paradoja de Moravec, y se explica porque nuestro cerebro hace que tales tareas parezcan fáciles al dedicar a ellas una enorme cantidad de hardware especializado (más de una cuarta parte de nuestro cerebro, de hecho).

Me encanta esta metáfora de Hans Moravec, y me he tomado la libertad de ilustrarla en la figura 2.2:

> Los ordenadores son máquinas universales, cuyo potencial abarca uniformemente una variedad ilimitada de tareas. Por su parte, las capacidades humanas son fuertes en áreas que durante mucho tiempo fueron importantes para la supervivencia, pero débiles en otras muy alejadas de aquellas. Imaginemos un «panorama de competencias humanas» con llanuras rotuladas como «aritmética» y «memorización», laderas como «demostración de teoremas» y «ajedrez» y elevados picos montañosos con nombres como «locomoción», «coordinación entre ojo y mano» e «interacción social». El avance del rendimiento de los ordenadores es como la lenta crecida de las aguas. Hace medio siglo, empezó a inundar las tierras bajas, expulsando a las calculadoras humanas y a los funcionarios de registros, pero permitiendo que la mayoría aún permaneciésemos secos. Ahora las aguas han llegado a las laderas, y los puestos fronterizos que tenemos allí están barajando retirarse. Nos sentimos a salvo en nuestros picos, pero, a la velocidad actual, estos también acabarán sumergidos dentro de otro medio siglo. Propongo que construyamos Arcas mientras se acerca ese día, y que nos vayamos preparando para una vida en alta mar.[5]

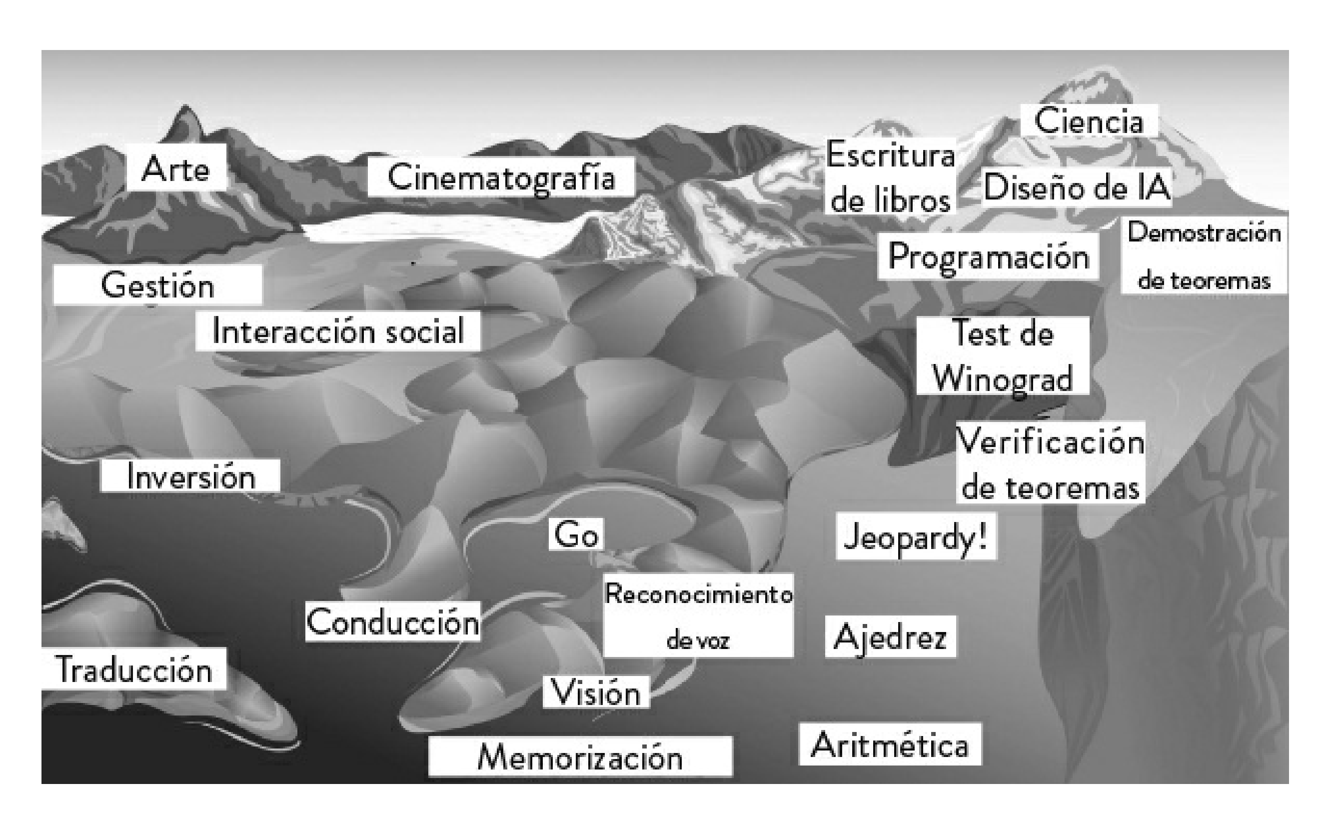


FIGURA 2.2. Ilustración del «panorama de competencias humanas» de Hans Moravec, en el que la elevación representa la dificultad para los ordenadores y el nivel creciente del mar representa lo que los ordenadores son capaces de hacer.

Durante las décadas transcurridas desde que escribió estas frases, el nivel del mar ha seguido subiendo constantemente, tal y como él predijo, como un calentamiento global sobreexcitado, y algunas de sus laderas (incluido el ajedrez) llevan ya tiempo sumergidas. Lo que vendrá a continuación y lo que debemos hacer al respecto es lo que nos ocupará en el resto del libro.

La subida del nivel del mar puede hacer que un día se llegue a un punto de inflexión a partir del cual se desencadene un cambio drástico. Esta altura crítica del nivel del mar corresponde al momento en que las máquinas sean capaces de diseñar IA. Antes de que se alcance este punto de inflexión, la subida del nivel de las aguas se debe a que los humanos mejoran las máquinas; después, la elevación puede deberse a que las máquinas mejoren las máquinas, potencialmente con mucha mayor rapidez de lo que los humanos podrían haberlo hecho, sumergiendo de forma vertiginosa toda la tierra. Esta es la fascinante y controvertida idea de la *singularidad*, que nos divertiremos explorando en el capítulo 4.

Como es bien sabido, Alan Turing, uno de los pioneros de la computación,

demostró que, si un ordenador es capaz de realizar un conjunto mínimo de operaciones, dándole el tiempo y la memoria suficientes, se puede programar para que haga cualquier cosa que cualquier otro ordenador pueda hacer. Las máquinas que superan este umbral crítico se denominan *ordenadores universales* (también conocidos como ordenadores universales de Turing); todos los teléfonos inteligentes y ordenadores portátiles actuales son, en este sentido, universales. Análogamente, me gusta ver el umbral crítico de inteligencia necesario para poder diseñar IA como el umbral de la *inteligencia universal*: si le damos el tiempo y los recursos suficientes, puede hacerse a sí misma capaz de lograr cualquier objetivo tan bien como cualquier otra entidad inteligente. Por ejemplo, si decide que quiere tener mejores habilidades sociales, mayor capacidad de previsión o de diseño de IA, puede adquirirlas. Si decide averiguar cómo construir una fábrica de robots, puede hacerlo. En otras palabras, la inteligencia universal es capaz de convertirse en vida 3.0.

La opinión generalizada entre los investigadores en inteligencia artificial es que, en última instancia, la inteligencia se reduce a información y computación, no a carne, sangre y átomos de carbono. En otras palabras, esto significa que no existe ninguna razón fundamental por la que las máquinas no puedan ser algún día al menos tan inteligentes como nosotros.

Pero qué son realmente la información y la computación, teniendo en cuenta que la física nos ha enseñado que, a un nivel fundamental, todas las cosas no son más que materia y energía moviéndose de acá para allá. ¿Cómo puede algo tan abstracto, intangible y etéreo como la información y la computación plasmarse en algo físico y tangible? En particular, ¿cómo pueden un montón de partículas tontas que se mueven de un sitio a otro según las leyes de la física exhibir un comportamiento que calificaríamos de inteligente?

Si cree que la respuesta a esta pregunta es evidente y le parece posible que las máquinas puedan llegar a ser tan inteligentes como los humanos en este siglo —por ejemplo, porque es usted un investigador en IA—, puede saltarse el resto de este capítulo y pasar directamente al capítulo 3. De lo contrario, le gustará saber que he escrito las tres secciones siguientes especialmente para usted.

Si afirmamos que un atlas contiene información sobre el mundo, lo que queremos decir es que existe una relación entre el estado del libro (en particular, las posiciones de ciertas moléculas que hacen que las letras e imágenes tengan los colores que tienen) y el estado del mundo (por ejemplo, la ubicación de los continentes). Si los continentes estuvieran situados en otros lugares, esas moléculas también ocuparían posiciones diferentes. Los humanos usamos una amplia variedad de dispositivos para almacenar información, desde libros y cerebros hasta discos duros, y todos ellos comparten esta propiedad: su estado puede guardar relación con (y, por lo tanto, informarnos sobre) el estado de otras cosas que nos interesan.

¿Qué propiedad física fundamental tienen todos ellos en común que los hace tan útiles como dispositivos de memoria, esto es, dispositivos para almacenar información? La respuesta es que todos *pueden encontrarse en muchos estados duraderos diferentes* (de duración suficiente para codificar la información hasta que se necesite). Como un ejemplo sencillo, imaginemos que colocamos una pelota sobre una superficie accidentada que tiene dieciséis valles distintos, como se muestra en la figura 2.3. Una vez que la pelota ha rodado hasta abajo y ha alcanzado el reposo, estará en uno de los dieciséis valles, por lo que podemos usar su posición como medio para recordar un número entre 1 y 16.

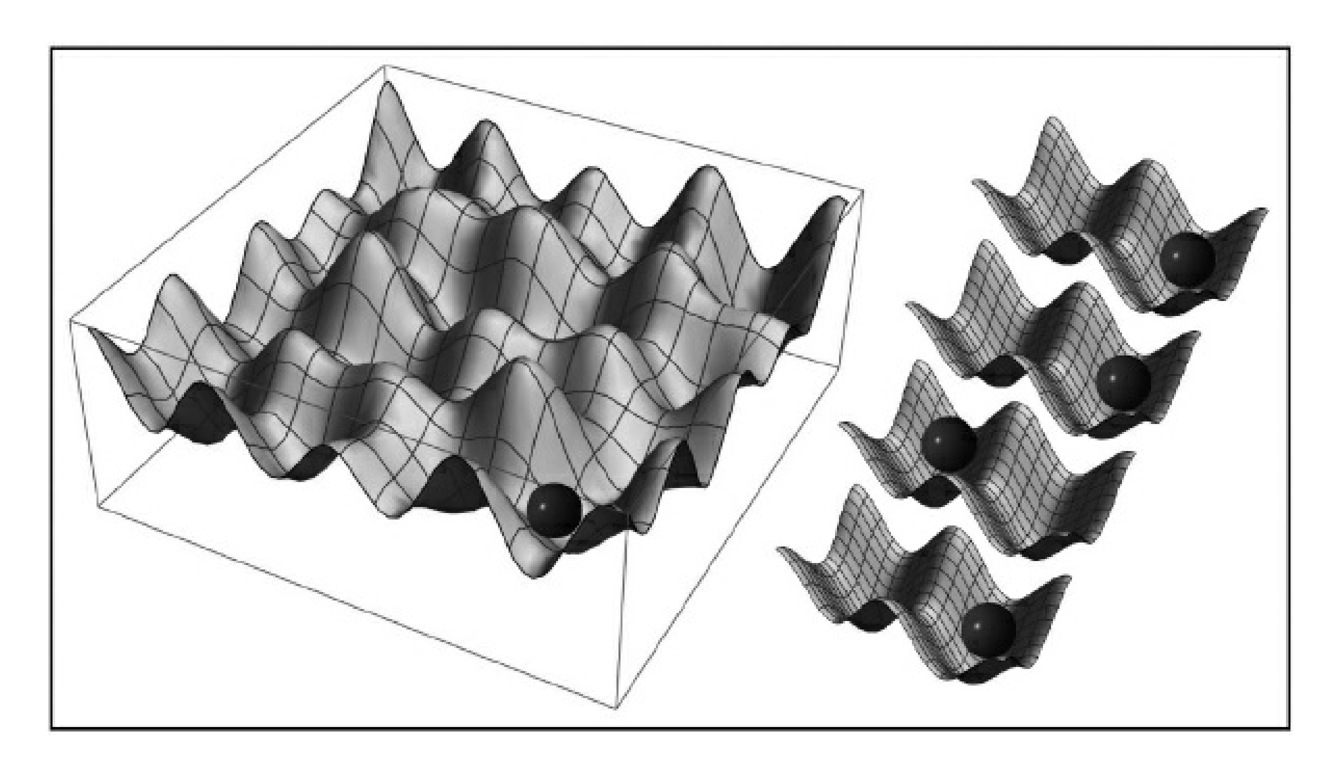


FIGURA 2.3. Un objeto físico constituye un dispositivo de memoria útil si puede encontrarse en muchos estados estables distintos. La bola en la imagen de la izquierda puede codificar los cuatro bits de información que especifican en cuál de los $2^4 = 16$ valles está. Juntas, las cuatro bolas de la derecha también codifican cuatro bits de información, uno cada una.

Este dispositivo de memoria es bastante robusto, porque es probable que la pelota, aunque sea ligeramente sacudida y perturbada por fuerzas externas, permanezca en el mismo valle en que la colocamos, por lo que seguiríamos sabiendo qué número se almacena. La razón por la que esta memoria es tan estable es que es necesaria más energía para elevar la pelota y sacarla de su valle de la producida normalmente por perturbaciones aleatorias. Esta misma idea puede dar lugar a memorias estables en forma mucho más general que una pelota móvil: la energía de un sistema físico complejo puede depender de toda clase de propiedades mecánicas, químicas, eléctricas y magnéticas, pero un estado será estable siempre que se necesite energía para sacar al sistema de ese estado que queremos que memorice. Este es el motivo por el que los sólidos tienen muchos estados duraderos, mientras que no sucede lo mismo en líquidos y gases: si grabamos un nombre en un anillo de oro, la información seguirá ahí años después porque se necesita una cantidad sustancial de energía para moldear el oro, mientras que si lo hacemos en la superficie de un estanque desaparecerá al instante, ya que la superficie del

agua cambia de forma sin dificultad.

El dispositivo de memoria más simple solo tiene dos estados posibles (figura 2.3). Podemos por lo tanto entender que codifica un dígito binario (un «bit»), esto es, un cero o un uno. La información almacenada en cualquier otro dispositivo de memoria más complicado puede guardarse de forma equivalente en varios bits: por ejemplo, considerados en conjunto, los cuatro bits que se representan en la figura 2.3 (a la derecha) pueden encontrarse en 2 × 2 × 2 × 2 = 16 estados diferentes (0000, 0001, 0010, 0011, …, 1111), por lo que, desde un punto de vista colectivo, tienen exactamente la misma capacidad de memoria que el sistema más complejo de 16 estados (izquierda). Podemos, por lo tanto, interpretar los bits como átomos de información: el pedazo mínimo de información que no puede seguir subdividiéndose y que puede combinarse para codificar cualquier información. Por ejemplo, acabo de escribir la palabra «palabra», y mi portátil la ha representado en su memoria como la secuencia de 7 números, 112 97 108 97 98 114 97, y ha almacenado cada uno de estos números como 8 bits (representa cada letra minúscula como un número que resulta de sumar 96 más su orden en el alfabeto). En cuanto pulso la tecla *p* en mi teclado, mi portátil muestra una imagen visual de una *p* en la pantalla, que también está representada por bits: 32 bits especifican el color de cada uno de los millones de píxeles de la pantalla.

Puesto que es fácil fabricar y trabajar con sistemas de dos estados, la mayoría de los ordenadores modernos almacenan su información en forma de bits, pero estos bits se plasman físicamente de muchas maneras distintas. En un DVD, cada bit corresponde a si hay o no una muesca microscópica en un determinado punto de la superficie plástica. En un disco duro, cada bit corresponde al hecho de que un punto de la superficie esté magnetizado en una de dos maneras posibles. En la memoria de trabajo de mi portátil, cada bit corresponde a las posiciones de determinados electrones, que determinan si un dispositivo denominado microcondensador está cargado o no. También hay tipos de bits fáciles de transportar, incluso a la velocidad de la luz: por ejemplo, en la fibra óptica por la que se transmite nuestro correo electrónico, cada bit corresponde a que un haz láser sea intenso o débil en un instante dado.

Los ingenieros prefieren codificar bits en sistemas que no sean solo estables y de los que sea fácil leer (como un anillo de oro), sino en los que

también resulte fácil escribir: alterar el estado de un disco duro requiere mucha menos energía que hacer una inscripción en oro. También prefieren sistemas con los que sea cómodo trabajar y que puedan producirse en masa. Pero, aparte de todo eso, no les importa cómo se representan los bits como objetos físicos, como tampoco nos importa a nosotros la mayoría de las veces, sencillamente porque da igual. Si le enviamos a un amigo por correo electrónico un documento para que lo imprima, la información se copiará en rápida sucesión para pasar de magnetizaciones en nuestro disco duro a cargas eléctricas en la memoria de trabajo en nuestro ordenador, a ondas de radio en nuestra red inalámbrica, voltajes en nuestro *router*, pulsos láser en una fibra óptica y, por último, moléculas en un pedazo de papel. Dicho de otro modo, *la información puede adquirir vida propia, independiente de su sustrato físico*. De hecho, normalmente lo único que nos interesa de la información es esta faceta independiente del sustrato: si nuestro amigo nos llama para comentar el documento que le enviamos, es probable que no quiera hablar de voltajes o moléculas. Este es el primer indicio que tenemos de que algo tan intangible como la inteligencia puede plasmarse en algo físico y tangible, y enseguida veremos cómo esta idea de la independencia respecto al sustrato es mucho más profunda, y abarca no solo a la información sino también a la computación y al aprendizaje.

Debido a esta independencia del sustrato, ha habido ingenieros ingeniosos capaces de sustituir de forma sucesiva los dispositivos de memoria en el interior de nuestros ordenadores por otros mucho mejores, basados en nuevas tecnologías, sin que fuese necesario hacer ninguna modificación en nuestro software. El resultado ha sido espectacular, como se ilustra en la figura 2.4: en las últimas seis décadas, el precio de la memoria de ordenador se ha reducido a la mitad aproximadamente cada dos años. Los discos duros son ahora cien millones de veces más baratos, y las memorias, mucho más rápidas y utilizadas para computación en lugar de para mero almacenamiento, son unos diez billones de veces más baratas. Asombroso. Si pudiésemos conseguir un descuento similar del 99,99999999999 % en todas nuestras compras, podríamos adquirir todos los inmuebles de Nueva York por unos diez centavos, y todo el oro que se ha extraído del suelo a lo largo de la historia por alrededor de un dólar.

Para muchos de nosotros, las espectaculares mejoras en la tecnología de las memorias están relacionadas con historias personales. Recuerdo con cariño

cuando trabajaba en una tienda de golosinas durante mis años de instituto para pagarme un ordenador con 16 kilobytes de memoria, y cuando desarrollé y vendí un procesador de texto con Magnus Bodin, un compañero de clase de entonces, y nos vimos obligados a escribirlo todo en código máquina ultracompacto para dejar memoria suficiente para las palabras que el programa debía procesar. Acostumbrado a que los disquetes tuviesen una capacidad de 70 kB, me dejaron pasmado los disquetes de 3,5 pulgadas, más pequeños y capaces de almacenar la asombrosa cifra de 1,44 MB, y en los que cabía un libro entero, y más tarde el primer disco duro que tuve, de 10 MB de capacidad (en el que apenas cabría una sola canción de las que nos descargamos ahora). Estos recuerdos de mi adolescencia me parecieron casi irreales hace unos días, cuando me gasté unos cien dólares en un disco duro con una capacidad 300.000 veces mayor.

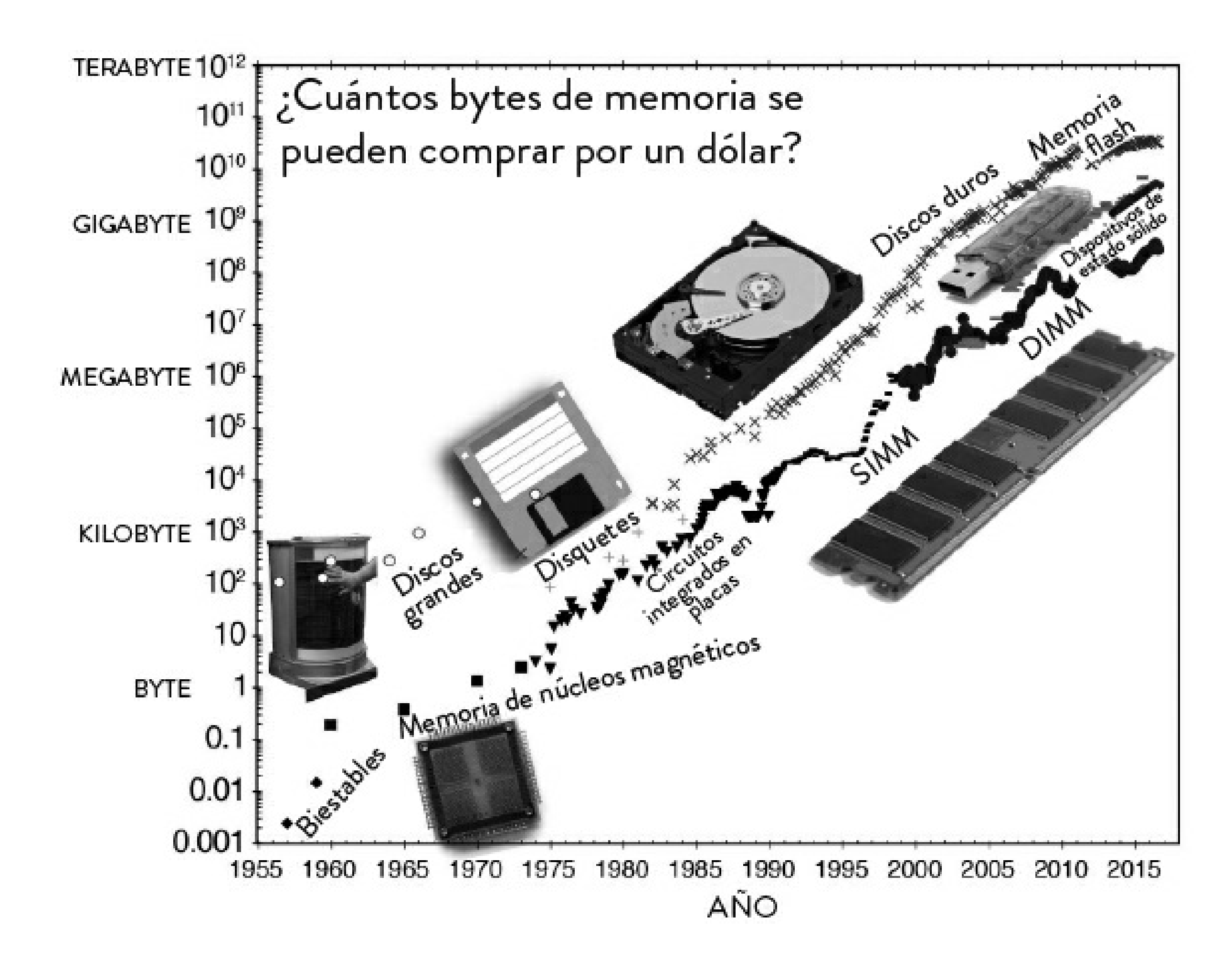


FIGURA 2.4. A lo largo de las seis décadas pasadas, el precio de la memoria de ordenador se ha reducido a la mitad aproximadamente cada dos años, lo que equivale a que sea unas mil veces más barata cada

veinte años. Un byte equivale a ocho bits. Datos cortesía de John McCallum, procedentes de <http://www.jcmit.net/memoryprice.htm>.

¿Y qué hay de los dispositivos de memoria que son producto de la evolución, en lugar de haber sido diseñados por los humanos? Los biólogos aún no saben cuál fue la primera forma de vida que copió sus propios planos a la siguiente generación, pero tuvo que haber sido bastante pequeña. Un equipo dirigido por Philipp Holliger en la Universidad de Cambridge creó en 2016 una molécula de ARN que codificaba 412 bits de información genética y era capaz de copiar cadenas de ARN más largas que ella misma, lo que reforzó la hipótesis del «mundo de ARN», según la cual la vida primitiva en la Tierra implicó pequeños fragmentos de ARN autorreplicantes. Hasta ahora, el dispositivo de memoria más pequeño que se sabe que es fruto de la evolución y se ha usado en la naturaleza es el genoma de la bacteria *Candidatus Carsonella ruddii*, que almacena alrededor de 40 kB, mientras que nuestro ADN contiene en torno a 1,6 GB, comparable al tamaño de una película descargada. Como se mencionó en el capítulo anterior, nuestro cerebro almacena mucha más información que nuestros genes: del orden de diez gigabytes eléctricamente (lo que especifica cuáles de nuestros cien mil millones de neuronas se activan en cada instante determinado) y cien terabytes química/biológicamente (lo que especifica la intensidad con la que distintas neuronas están conectadas entre sí mediante sinapsis). Si comparamos estas cifras con las de las memorias artificiales, vemos que los mejores ordenadores del mundo superan ya la capacidad de recordar de cualquier sistema biológico, y a un coste que disminuye rápidamente y que en 2016 era de unos pocos miles de dólares.

La memoria en nuestro cerebro funciona de manera muy diferente de cómo lo hace una memoria de ordenador, no solo en lo que se refiere a cómo está construida, sino también a cómo se usa. Mientras que recuperamos los datos de un ordenador o disco duro especificando dónde están almacenados, los recuerdos almacenados en nuestro cerebro se recuperan especificando sobre qué tratan. Cada grupo de bits en la memoria de nuestro ordenador tiene una dirección numérica y, para recuperar un dato, el ordenador especifica en qué dirección mirar, como si yo le dijese: «Vaya a mi librería, coja el quinto libro desde la derecha en la balda superior y dígame lo que pone en la página 314». Por el contrario, extraemos información de nuestro cerebro de forma similar a

cómo se obtiene de un buscador: especificamos parte de la información o algo relacionado con ella, y esta aparece. Si le digo a usted «ser o no ser», o si lo busco en Google, es muy posible que la respuesta sea «Ser o no ser, esa es la cuestión». De hecho, probablemente también funcionaría si usase otra parte de la cita o revolviese un poco las cosas. Estos sistemas de memoria se denominan autoasociativos, ya que recuerdan por asociación, en lugar de hacerlo por dirección.

En un famoso artículo de 1982, el físico John Hopfield demostró cómo una red de neuronas interconectadas podría funcionar como una memoria autoasociativa. La idea básica me parece muy hermosa, y vale para cualquier sistema físico con varios estados estables. Por ejemplo, consideremos una bola en una superficie con dos valles, como el sistema de un bit en la figura 2.3, y hagamos que la superficie tenga una forma tal que las coordenadas x de los dos mínimos donde la bola puede acabar reposando sean $x = \sqrt{2} \approx 1{,}41421$ y $x = \pi \approx 3{,}14159$, respectivamente. Si tan solo recordamos que π tiene un valor próximo a 3, colocamos la bola en $x = 3$ y observamos cómo nos muestra un valor más preciso de π al rodar hasta el mínimo más cercano. Hopfield se dio cuenta de que una red neuronal compleja constituía un paisaje análogo con muchísimos mínimos de energía en los que el sistema podía establecerse, y más tarde se demostró que se pueden llegar a apretujar hasta 138 recuerdos diferentes por cada mil neuronas sin causar una gran confusión.

¿QUÉ ES LA COMPUTACIÓN?

Hemos visto cómo un objeto físico puede recordar información, pero ¿cómo puede computar algo?

Una computación es una transformación desde un estado de memoria hasta otro. Dicho de otro modo, una computación toma información y la transforma, implementando lo que los matemáticos llaman una *función*. Imagino una función como una picadora de información, como se ilustra en la figura 2.5: se introduce la información por la parte superior, se hace girar la manivela y se obtiene información procesada por la parte inferior; y se puede repetir este proceso tantas veces como se quiera con distinto material de entrada. Este procesamiento de información es determinista en el sentido de

que si se repite con la misma información de entrada se obtiene siempre la misma información a la salida.

Aunque pueda parecer engañosamente sencillo, este concepto de función es algo muy común. Algunas funciones son triviales, como la denominada NOT, que toma un único bit y devuelve el inverso, transformando un cero en un uno, y viceversa. Las funciones que estudiamos en el colegio corresponden habitualmente a botones de una calculadora de bolsillo y en ellas se introduce uno o más números y dan como resultado uno solo (por ejemplo, la función x^2 se limita a tomar un número y lo devuelve multiplicado por sí mismo). Otras funciones pueden ser muy complicadas. Por ejemplo, si disponemos de una que toma los bits que representan una posición arbitraria en el ajedrez y devuelve los bits que representan el mejor movimiento posible a partir de dicha posición, podemos utilizarla para ganar el Campeonato Mundial de Ajedrez por Ordenador. Si disponemos de una función que tome todos los datos financieros del mundo y señale las acciones que es mejor comprar, rápidamente nos haremos muy ricos. Muchos investigadores en IA dedican sus carreras a averiguar cómo implementar determinadas funciones. Por ejemplo, el objetivo de la investigación en traducción automática es implementar una función en la que, si se introdujesen los bits que representan un texto en un idioma, devolviese los bits que representan ese mismo texto en otro idioma, y el objetivo de la investigación en subtitulación automática es tomar los bits que representan una imagen y devolver los bits que representan un texto que la describe (figura 2.5, derecha).

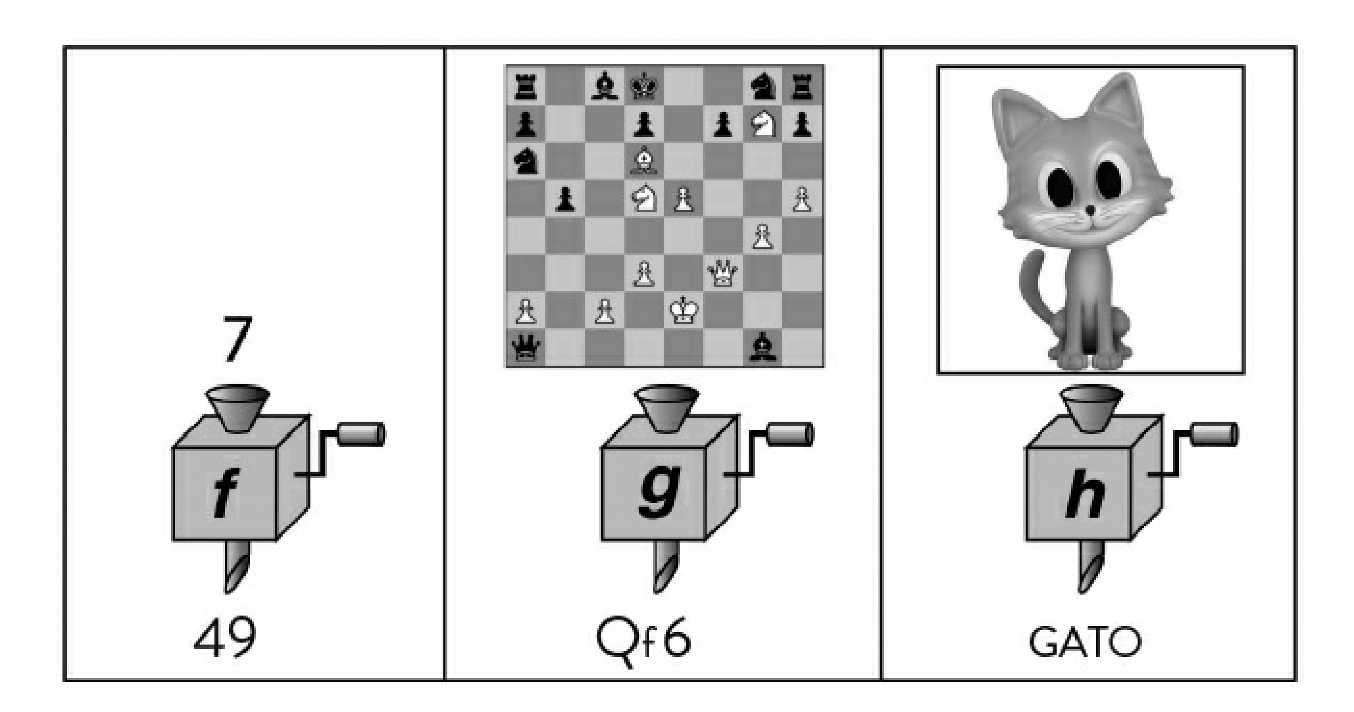


FIGURA 2.5. Una *computación* toma información y la transforma, implementando lo que los matemáticos llaman una *función*. La función *f* (izquierda) toma los bits que representan un número y computa el cuadrado del mismo. La función *g* (centro) toma los bits que representan una posición de ajedrez y computa el mejor movimiento para las blancas. La función *h* (derecha) toma los bits que representan una imagen y computa un texto que la describa.

Dicho de otro modo, si podemos implementar funciones altamente complejas, entonces podemos construir una máquina inteligente que sea capaz de alcanzar objetivos muy complejos. Esto centra más nuestra pregunta sobre cómo puede la materia ser inteligente: en particular, ¿cómo puede una masa de materia aparentemente tonta computar una función complicada?

En lugar de permanecer inmóvil como un anillo de oro u otro dispositivo de memoria estático, debe exhibir una dinámica compleja tal que su estado futuro dependa de alguna manera complicada (y, con suerte, controlable/programable) del estado actual. La disposición de sus átomos debe ser menos ordenada que la de un sólido rígido, donde nada interesante cambia, pero más que la de un líquido o un gas. Lo que queremos es que, si colocamos el sistema en un estado que codifica la información de entrada, nos permita dejarlo evolucionar de acuerdo con las leyes de la física durante cierto periodo de tiempo, y a continuación interpretar el estado final resultante como la información de salida, y que dicha salida sea la función deseada de la entrada. Si es así, podemos decir que nuestro sistema computa

esa función.

Como primer ejemplo de esta idea, exploremos cómo podemos construir una función muy simple (pero también muy importante) llamada *puerta* NAND(6) a partir de materia tonta normal de toda la vida. Esta función recibe dos bits y devuelve un bit: devuelve 0 si ambas entradas son 1; en todos los demás casos, devuelve 1. Si conectamos dos interruptores en serie con una batería y un electroimán, este último solo se activará si el primer interruptor y el segundo están cerrados («encendidos»). Coloquemos un tercer interruptor bajo el electroimán, como se ilustra en la figura 2.6, de tal manera que, cuando el imán esté activado, su atracción provoque que el interruptor se abra. Si interpretamos los dos primeros interruptores como los bits de entrada y el tercero como el bit de salida (con 0 = interruptor abierto y 1 = interruptor cerrado), entonces tenemos una puerta NAND: el tercer interruptor se abre solo si los dos primeros están cerrados. Hay muchas otras maneras de construir puertas NAND que resulten más prácticas; por ejemplo, usando transistores, tal y como se muestra en la figura 2.6. En los ordenadores actuales, las puertas NAND suelen estar construidas a partir de transistores y otros componentes que pueden grabarse automáticamente en obleas de silicio.

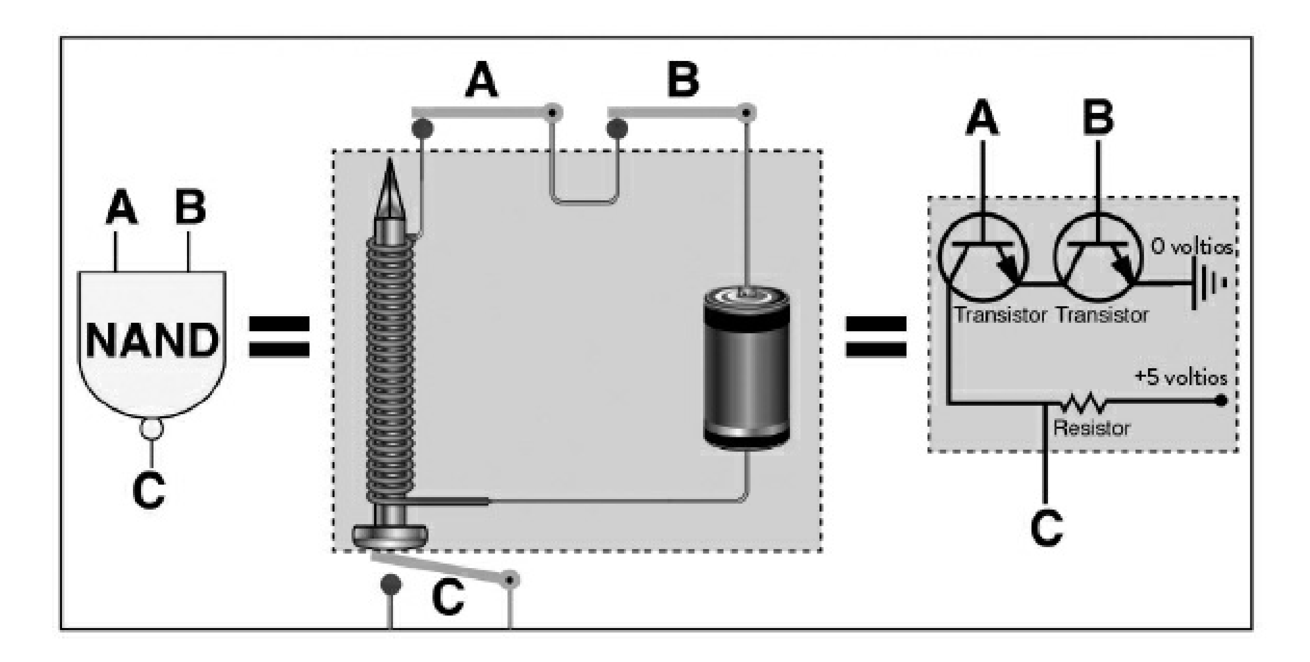


FIGURA 2.6. Una puerta NAND toma como entradas dos bits, A y B, y computa como salida un bit C, de acuerdo con la regla de que C = 0 si A = B = 1 y C = 1 en los demás casos. Muchos sistemas físicos

pueden usarse como puertas NAND. En el ejemplo del centro, los interruptores se interpretan como bits (0 = abierto, 1 = cerrado) y, cuando los interruptores A y B están cerrados, un electroimán abre el interruptor C. En el ejemplo de la derecha, los voltajes (potenciales eléctricos) se interpretan como bits (1 = cinco voltios, 0 = cero voltios) y, cuando los cables A y B están a cinco voltios, los dos transistores conducen la electricidad, y el cable C cae hasta un potencial de aproximadamente cero voltios.

Hay un notable teorema en informática que dice que las puertas NAND son universales, lo que significa que se puede implementar cualquier función bien definida simplemente conectando entre sí puertas NAND.(7) Así pues, si podemos construir una cantidad suficiente de puertas NAND, también podremos construir un dispositivo capaz de computar cualquier cosa. Si quiere hacerse una idea de cómo funciona esto, en la figura 2.7 he representado cómo multiplicar números usando tan solo puertas NAND.

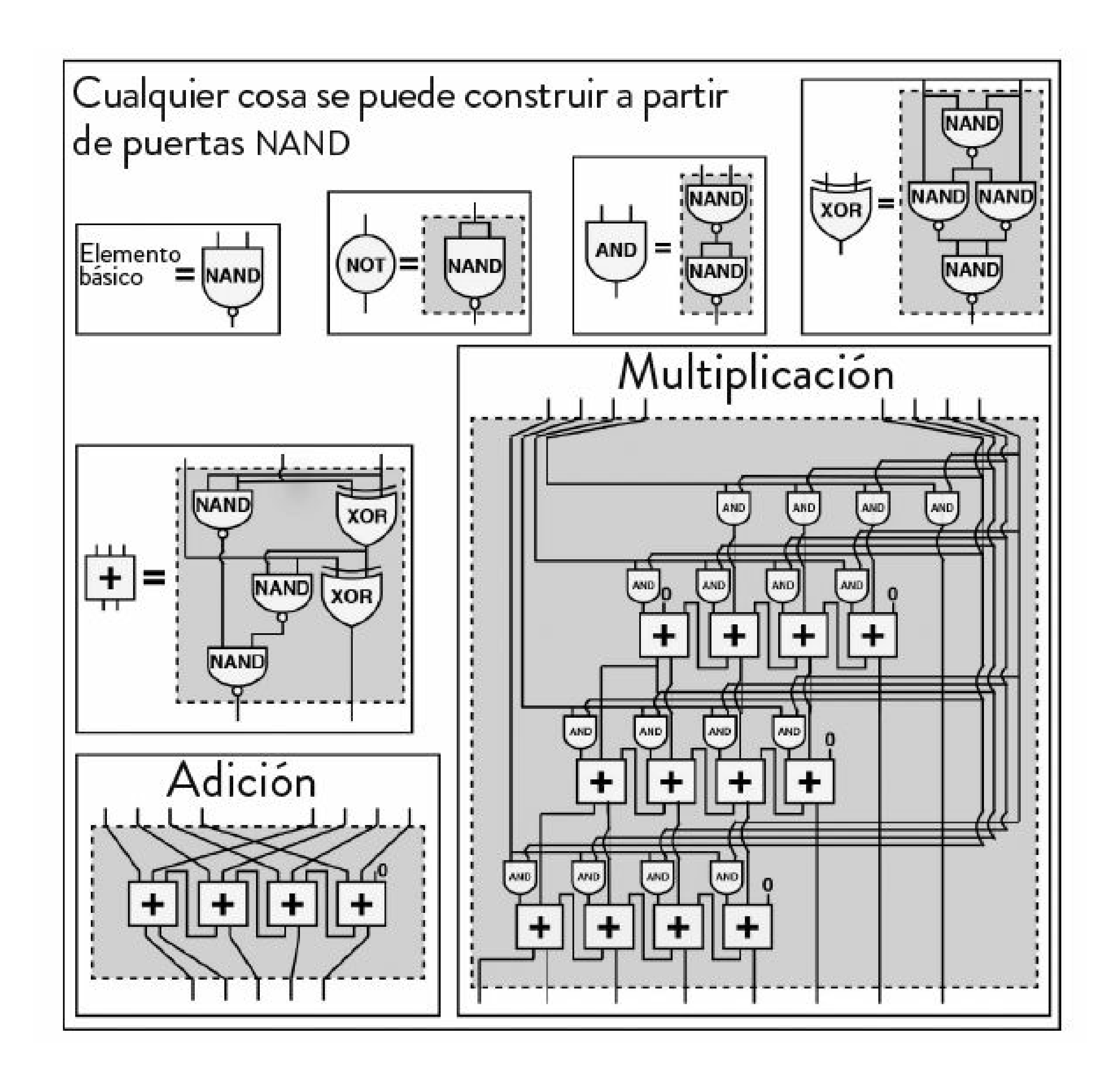


FIGURA 2.7. Cualquier computación bien definida puede realizarse combinando solo puertas NAND de forma ingeniosa. Por ejemplo, los módulos de adición y multiplicación que se representan en esta figura toman dos números binarios representados mediante 4 bits y devuelven un número binario presentado mediante 5 y 8 bits, respectivamente. Los módulos más pequeños, NOT, AND, XOR y + (que suma tres bits distintos para dar un número binario de 2 bits), están a su vez compuestos de puertas NAND. Entender del todo esta figura es muy difícil e innecesario para seguir lo que se explica en el resto del libro; la incluyo aquí para ilustrar la idea de universalidad, y para satisfacer al *geek* que llevo dentro.

Norman Margolus y Tommaso Toffoli, investigadores en el MIT, acuñaron el nombre *computronio* para referirse a cualquier sustancia capaz de realizar computaciones arbitrarias. Acabamos de ver que fabricar computronio no es particularmente difícil: la sustancia solo debe ser susceptible de implementar

puertas NAND conectadas entre sí de la manera que se quiera. De hecho, existen multitud de otros tipos de computronio. Una variante sencilla que también funciona se obtiene al sustituir las puertas NAND por puertas NOR, que devuelven 1 solo cuando ambas entradas son 0. En la siguiente sección, veremos las redes neuronales, que también implementan computaciones arbitrarias, esto es, actúan como computronio. El científico y emprendedor Stephen Wolfram ha demostrado que lo mismo vale también para dispositivos simples llamados «autómatas celulares», que actualizan repetidamente el valor de unos bits basándose en lo que hacen los bits vecinos. Ya en 1936, Alan Turing, pionero de la computación, demostró en un artículo seminal que una máquina sencilla (conocida como «máquina universal de Turing»), capaz de manipular símbolos en una cinta de papel, podía también implementar computaciones arbitrarias. Resumiendo: no solo es posible que la materia implemente cualquier computación bien definida, sino que puede hacerlo en infinidad de maneras.

Como ya se ha mencionado, Turing también demostró algo todavía más profundo en ese artículo suyo de 1936: que si un tipo de ordenador puede realizar un determinado conjunto mínimo de operaciones entonces es *universal* en el sentido de que, dotado de los recursos suficientes, puede hacer cualquier cosa que otro ordenador sea capaz de hacer. Demostró que su máquina de Turing era universal y, conectando más directamente con la física, acabamos de ver que esta familia de ordenadores universales también incluye objetos tan diversos como una red de puertas NAND o una red de neuronas interconectadas. De hecho, Stephen Wolfram ha sostenido que la mayoría de los sistemas físicos no triviales, desde sistemas meteorológicos hasta cerebros, serían ordenadores universales si pudiesen hacerse arbitrariamente grandes y duraderos.

El hecho de que exactamente la misma computación pueda realizarse en cualquier ordenador universal significa que la *computación es independiente del sustrato*, de la misma manera en que lo es la información: puede tomar vida propia, independiente de su sustrato físico. Así pues, si fuésemos un personaje consciente y superinteligente en un videojuego del futuro, no tendríamos manera de saber si nos estamos ejecutando en un ordenador de escritorio con Windows, en un portátil con Mac OS o en un teléfono con Android, ya que seríamos independientes del sustrato. Tampoco podríamos saber qué tipos de transistores usa el microprocesador.

Me di cuenta de este concepto fundamental de independencia respecto del sustrato porque en física existen muchos ejemplos hermosos. Las ondas, por ejemplo: tienen propiedades como la velocidad, la longitud de onda y la frecuencia, y los físicos podemos estudiar las ecuaciones que satisfacen sin necesidad de saber siquiera en qué sustancia en particular se propagan. Cuando oímos algo, estamos detectando ondas sonoras provocadas por la agitación de las moléculas en la mezcla de gases que llamamos «aire», y podemos determinar ciertas características de estas ondas —cómo disminuye su intensidad con el cuadrado de la distancia, cómo se curvan cuando atraviesan puertas abiertas o cómo rebotan en las paredes dando lugar a ecos— sin saber de qué está compuesto el aire. De hecho, ni siquiera necesitamos saber que está formado por moléculas: podemos ignorar todos los detalles relativos al oxígeno, al nitrógeno, al dióxido de carbono y demás, porque la única propiedad del sustrato de la onda que importa y aparece en la famosa ecuación de onda es un solo número que podemos medir: la velocidad de propagación de la onda, que en este caso es de unos 300 metros por segundo. De hecho, esta ecuación de onda que les expliqué a mis alumnos en el MIT en una asignatura la primavera pasada fue descubierta y usada con gran éxito mucho antes de que los físicos hubiesen siquiera demostrado que los átomos y las moléculas existen.

El ejemplo de las ondas ilustra tres ideas importantes. En primer lugar, que la independencia respecto del sustrato no significa que este no sea necesario, sino que la mayoría de sus detalles no son relevantes. Obviamente, no podemos tener ondas sonoras en un gas si no hay gas, pero cualquier gas, sea el que sea, nos valdría. De forma análoga, es evidente que no podemos tener computación sin materia, pero cualquier tipo de materia valdrá siempre que esta pueda organizarse en puertas NAND, neuronas conectadas o algún otro elemento básico que haga posible la computación universal. Segundo, el fenómeno independiente del sustrato adquiere vida propia, independiente de su sustrato. Una onda puede propagarse a través de un lago, aunque ninguna de sus moléculas lo hagan (se limitan básicamente a oscilar arriba y abajo, como los espectadores que hacen la ola en un estadio deportivo). Tercero, a menudo solo nos interesa el hecho de que un fenómeno sea independiente del sustrato: a un surfista le importa más la posición y altura de una ola que el detalle de su composición molecular. Vimos cómo esto era cierto para la información, y también lo es para la computación: si dos programadores

tratan conjuntamente de localizar un error en su código, es probable que no hablen de transistores.

Hemos llegado a una respuesta a nuestra pregunta inicial sobre cómo algo físicamente tangible puede dar lugar a algo que parece tan intangible, abstracto y etéreo como la inteligencia: resulta tan poco físico porque es independiente del sustrato, y adquiere una vida propia que no refleja los detalles físicos ni depende de ellos. En resumen: la computación es una pauta en la disposición espacio-temporal de las partículas, y lo realmente importante no son las partículas sino la pauta.

En otras palabras, el hardware es la materia, y el software, la pauta. Esta independencia respecto del sustrato que caracteriza a la computación implica que la IA es posible: la inteligencia no necesita carne, sangre o átomos de carbono.

Debido a esta independencia del sustrato, astutos ingenieros han sido capaces de sustituir una y otra vez las tecnologías internas de nuestros ordenadores por otras muchísimo mejores sin tener que modificar el software. Los resultados han sido tan espectaculares como para los dispositivos de memoria. Como se ilustra en la figura 2.8, el coste de la computación se reduce a la mitad aproximadamente cada dos años, y esta tendencia persiste desde hace más de un siglo, lo que ha hecho que el coste de un ordenador sea nada menos que un millón de millones de millones (10^{18}) de veces menor que cuando nacieron mis abuelas. Si todo fuese un millón de millones de millones más barato, con la centésima parte de un centavo podrían comprarse todos los bienes y servicios producidos en la Tierra este año. Esta espectacular reducción de los costes es, qué duda cabe, uno de los motivos clave por los que la computación está hoy en día en todas partes, y se ha extendido desde las instalaciones de computación de antaño, grandes como edificios, hasta nuestros hogares, coches y bolsillos, e incluso aparece en lugares tan inesperados como nuestras zapatillas deportivas.

¿Por qué la tecnología dobla su potencia a intervalos regulares y exhibe lo que los matemáticos llaman un crecimiento exponencial? De hecho, ¿por qué sucede no solo en lo que respecta a la miniaturización de los transistores (una tendencia que se conoce como *ley de Moore*), sino también de manera más amplia para la computación en su conjunto (figura 2.8), para la memoria (figura 2.4) y para multitud de otras tecnologías que van desde la

secuenciación del genoma hasta las imágenes del cerebro? Ray Kurzweil se refiere a este fenómeno de duplicación persistente como «la ley de rendimientos acelerados».

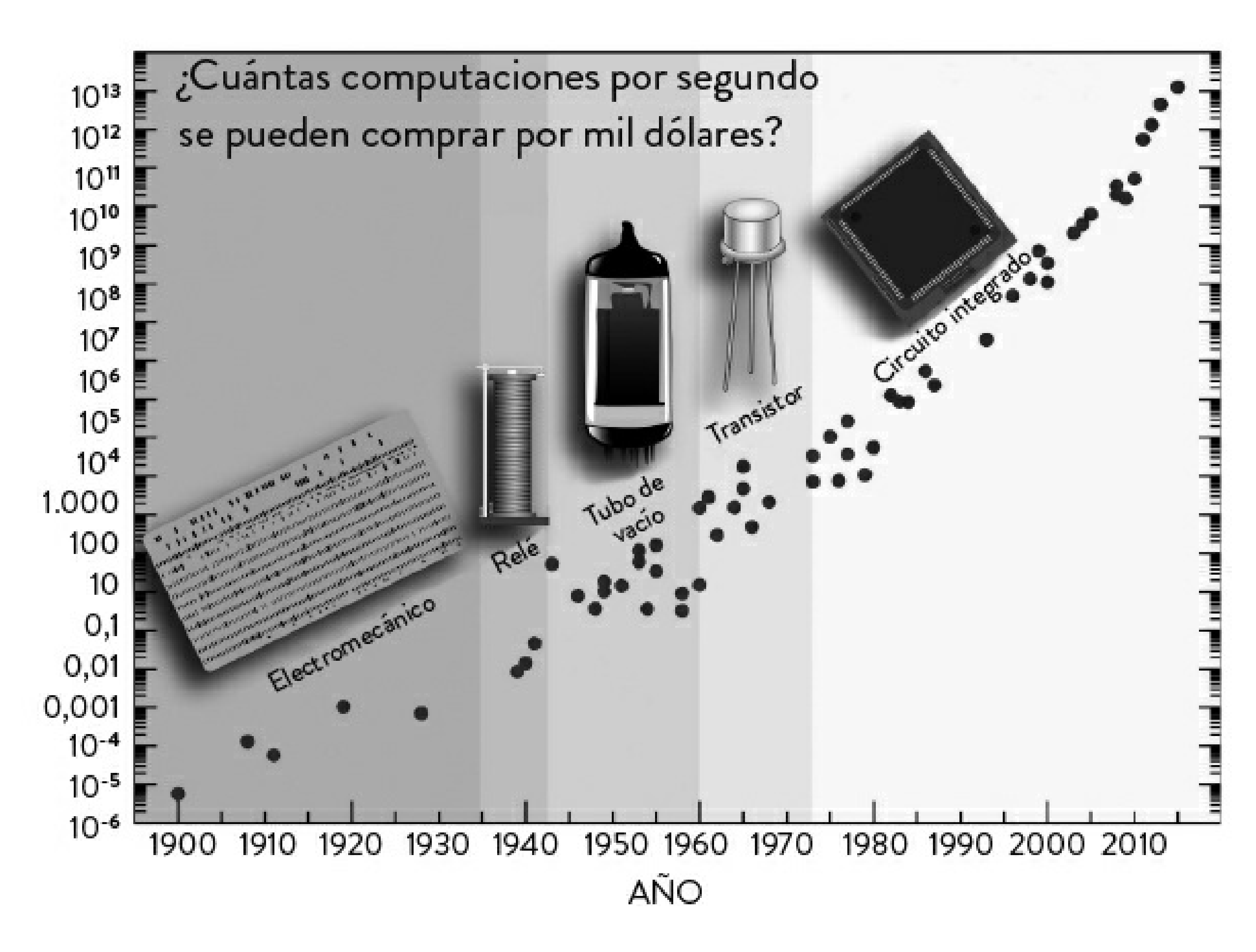


FIGURA 2.8. Desde 1900, el coste de la computación se ha reducido a la mitad aproximadamente cada dos años. La gráfica muestra la potencia de computación medida en operaciones de punto flotante por segundo (FLOPS) que pueden comprarse por mil dólares.[6] La computación concreta que define una operación de punto flotante corresponde a unas 10^5 operaciones lógicas elementales, como inversiones de bits o evaluaciones NAND.

Todos los ejemplos de duplicación persistente de los que tengo conocimiento en la naturaleza tienen la misma causa fundamental, y este caso tecnológico no es una excepción: cada paso origina el siguiente. Por ejemplo, nosotros experimentamos un crecimiento exponencial después de nuestra concepción: cada una de nuestras células se dividió y, con un ritmo más o menos diario, dio lugar a otras dos células, lo que hizo que nuestro número total de células aumentase día a día de 1 a 2, 4. 8, 16, y así sucesivamente.

Según la teoría científica más popular del origen del cosmos, conocida como *inflación*, en un determinado momento nuestro universo recién nacido creció de forma exponencial, igual que nosotros, y duplicó su tamaño una y otra vez a intervalos regulares, hasta que una mota mucho más pequeña y ligera que un átomo llegó a ser más enorme que todas las galaxias que hemos visto jamás con nuestros telescopios. De nuevo, la causa fue un proceso por el cual cada paso de duplicación provocaba el siguiente. Así es también como progresa la tecnología: una vez que una tecnología se vuelve el doble de potente, a menudo puede utilizarse para diseñar y construir otra que sea a su vez el doble de potente, dando pie a una capacidad de duplicación en cadena a tenor de la ley de Moore.

Algo que sucede con tanta regularidad como la duplicación de nuestra potencia tecnológica es la aparición de voces que afirman que la duplicación está llegando a su fin. Sí, claro que la ley de Moore tendrá un final (lo que significa que existe un límite físico para lo pequeños que pueden hacerse los transistores), pero hay quien asume erróneamente que la ley de Moore equivale a la duplicación persistente de nuestra potencia tecnológica. Por el contrario, Ray Kurzweil señala que la ley de Moore no se apoya en el primero sino en el quinto paradigma tecnológico para generar un crecimiento exponencial en la computación, como se ilustra en la figura 2.8: cuando ya no es posible mejorar una tecnología, la reemplazamos por otra mejor. Cuando no pudimos seguir reduciendo el tamaño de los tubos de vacío, los sustituimos por transistores, y después por circuitos integrados, en los que los electrones se mueven en dos dimensiones. Cuando esta tecnología alcance su límite, hay muchas alternativas que podemos probar; por ejemplo, utilizar circuitos tridimensionales y algo que no sean electrones para que cumplan nuestras órdenes.

Nadie sabe a ciencia cierta cuál será el próximo sustrato computacional de éxito, pero sí sabemos que aún no estamos ni remotamente cerca del límite que imponen las leyes físicas. Mi colega en el MIT Seth Lloyd ha calculado cuál es este límite fundamental y, como veremos en detalle en el capítulo 6, este límite es nada menos que 33 órdenes de magnitud (10^{33} veces) superior al poder de computación que la tecnología punta actual es capaz de extraer de un pedazo de materia. De manera que, aunque siguiésemos doblando la potencia de nuestros ordenadores cada dos años, aún se tardarían más de dos

siglos en alcanzar esa frontera última.

Aunque todos los ordenadores universales son capaces de realizar las mismas computaciones, algunos son más eficientes que otros. Por ejemplo, una computación que requiera millones de multiplicaciones no precisa de millones de módulos de multiplicación separados construidos a partir de transistores distintos, como en la figura 2.6: solo necesita uno de estos módulos, ya que puede usarlo sucesivamente muchas veces con las entradas apropiadas. Para mejorar la eficiencia, la mayoría de los ordenadores modernos utilizan un paradigma en el que las computaciones se dividen en múltiples pasos temporales, durante los cuales la información va y viene entre los módulos de memoria y los módulos de computación. Esta arquitectura computacional fue desarrollada entre 1935 y 1945 por pioneros de la computación, entre los que estaban Alan Turing, Konrad Zuse, Presper Eckert, John Mauchly y John von Neumann. Más específicamente, la memoria del ordenador almacena tanto datos como software (un programa, esto es, una lista de instrucciones sobre lo que hacer con los datos). En cada paso temporal, una unidad central de procesamiento (CPU) ejecuta la siguiente instrucción del programa, que especifica qué función simple hay que aplicar a alguna parte de los datos. La parte del ordenador que lleva la cuenta de qué hay que hacer a continuación no es más que otra parte de su memoria, llamada *contador de programa*, que almacena el número de línea actual en el programa. Para pasar a la siguiente instrucción, solo hay que sumar uno al contador de programa. Para saltar a otra línea del programa, simplemente se copia ese número de línea en el contador de programa (así es como se implementan las llamadas sentencias «if» y los bucles).

Los ordenadores actuales suelen incrementar su velocidad recurriendo al *procesamiento en paralelo*, que de forma ingeniosa deshace parte de esta reutilización de los módulos: si una computación puede dividirse en partes que pueden llevarse a cabo en paralelo (porque la entrada de una parte no necesita de la salida de otra), entonces dichas partes pueden computarse simultáneamente por distintas partes del hardware.

El ordenador paralelo por antonomasia es el *ordenador cuántico*. David Deutsch, pionero de la computación cuántica, defiende la polémica idea de que «los ordenadores cuánticos comparten información con una enorme cantidad de versiones de sí mismos existentes en el multiverso», y pueden obtener resultados más rápidamente aquí en nuestro universo gracias, en

algún sentido, a la ayuda que reciben de esas otras versiones.[7] Aún no sabemos si en las próximas décadas se podrá construir un ordenador cuántico competitivo desde un punto de vista comercial, porque eso depende tanto de si la física cuántica funciona como creemos, como de nuestra capacidad de superar desalentadores retos técnicos. Sin embargo, empresas y gobiernos de todo el mundo están apostando decenas de millones de dólares cada año a que esa posibilidad existe. Aunque los ordenadores cuánticos no pueden acelerar las computaciones corrientes, se han desarrollado ingeniosos algoritmos que podrían acelerar de forma espectacular determinados tipos de cálculos, como el craqueo de criptosistemas o el entrenamiento de redes neuronales. Un ordenador cuántico también podría simular eficientemente el comportamiento de sistemas mecano-cuánticos, incluidos átomos, moléculas y nuevos materiales, reemplazando así las mediciones en los laboratorios de química, de la misma manera en que las simulaciones en ordenadores tradicionales han sustituido las mediciones en los túneles de viento.

¿QUÉ ES EL APRENDIZAJE?

Aunque una calculadora de bolsillo puede machacarme en una competición de aritmética, nunca podrá mejorar su velocidad o su precisión, por mucho que practique: no aprende. Por ejemplo, cada vez que pulso su botón de raíz cuadrada, computa exactamente la misma función de la misma manera. De forma análoga, el primer programa de ordenador que me ganó al ajedrez nunca aprendió de sus errores, sino que se limitó a implementar una función que su ingenioso programador había diseñado para computar un buen movimiento siguiente. Por el contrario, cuando Magnus Carlsen perdió su primera partida de ajedrez a los cinco años, empezó un proceso de aprendizaje que lo llevó a convertirse en campeón del mundo de ajedrez dieciocho años más tarde.

La capacidad de aprender es sin duda el aspecto más fascinante de la inteligencia general. Ya hemos visto cómo un pedazo aparentemente tonto de materia puede recordar y computar, pero ¿cómo puede aprender? Hemos visto que encontrar la respuesta a una pregunta difícil corresponde a computar una función, y que, si está bien organizada, la materia es capaz de calcular cualquier función computable. Cuando los humanos creamos las

calculadoras de bolsillo y los programas de ajedrez, fuimos nosotros quienes organizamos la materia. Para que esta aprenda, debe reorganizarse *por sí sola* para mejorar de forma progresiva su capacidad de computar la función en cuestión, y hacerlo obedeciendo las leyes de la física.

Para clarificar el proceso de aprendizaje, consideremos primero cómo un sistema físico muy simple puede aprender los dígitos de π y otros números. Antes vimos cómo una superficie con muchos valles (véase la figura 2.3) puede utilizarse como un dispositivo de memoria: por ejemplo, si el fondo de uno de los valles está en la posición $x = \pi \approx 3{,}14159$ y no hay otros valles cerca, podemos colocar la bola en $x = 3$ y observar cómo el sistema computa los decimales que faltan al dejar que la bola ruede ladera abajo hasta el fondo del valle. Ahora, supongamos que la superficie está hecha de arcilla blanda y comienza siendo completamente plana, como una pizarra en blanco. Si unos entusiastas de las matemáticas colocan una y otra vez la bola en las posiciones de cada uno de sus números favoritos, la gravedad irá creando valles en esas ubicaciones, tras lo cual la superficie de arcilla podrá usarse para rememorar esos recuerdos almacenados. Dicho de otro modo, la superficie de arcilla ha aprendido a computar los dígitos de números como π.

Otros sistemas físicos, como los cerebros, pueden aprender con mucha más eficiencia basándose en la misma idea. John Hopfield demostró que su ya mencionada red de neuronas interconectadas puede aprender de manera análoga: si se la coloca repetidamente en determinados estados, aprenderá de forma gradual dichos estados y volverá a ellos desde cualquier otro cercano. Si hemos visto muchas veces a cada uno de nuestros familiares, cualquier cosa relacionada con ellos puede disparar los recuerdos de su aspecto.

Las redes neuronales han transformado la inteligencia tanto biológica como artificial, y no hace mucho han empezado a dominar el subcampo de la IA conocido como *aprendizaje automático* (el estudio de algoritmos que mejoran por medio de la experiencia). Antes de profundizar más en cómo pueden aprender tales redes, tratemos de entender primero cómo pueden computar. Una red neuronal es un grupo de neuronas interconectadas que son capaces de influirse las unas a las otras sobre su comportamiento. Nuestro cerebro contiene aproximadamente tantas neuronas como estrellas hay en nuestra galaxia: del orden de cien mil millones. En promedio, cada de estas neuronas está conectada con alrededor de otras mil a través de uniones llamadas *sinapsis*, y son las intensidades de estos cien billones de conexiones

sinápticas las que codifican la mayoría de la información en nuestro cerebro.

Podemos dibujar esquemáticamente una red neuronal como un conjunto de puntos que representan las neuronas, conectados mediante líneas que representan las sinapsis (véase la figura 2.9). Las neuronas reales son dispositivos electroquímicos muy complejos que no se parecen en nada a esta representación esquemática: tienen distintas partes, con nombres como axones y dendritas, existen muchos tipos diferentes de neuronas que operan de muy diversas maneras, y los detalles precisos de cómo y cuándo la actividad de una de las neuronas afecta a otras es aún objeto activo de estudio. Sin embargo, los investigadores en IA han demostrado que las redes neuronales pueden alcanzar un rendimiento de nivel humano en muchas tareas extraordinariamente complejas, incluso si se ignoran las complejidades y se sustituyen las neuronas biológicas reales por otras simuladas muy sencillas, que son todas ellas idénticas y obedecen las mismas reglas, muy sencillas también. Hoy en día, el modelo más popular para una de estas *redes neuronales artificiales* representa el estado de cada neurona mediante un único número, y la intensidad de cada sinapsis mediante otro solo número. En este modelo, cada neurona actualiza su estado a intervalos de tiempo regulares calculando un promedio de todas las entradas de las neuronas con las que está conectada, sopesándolas en función de las intensidades sinápticas, añadiendo de forma opcional una constante, y a continuación aplicando al resultado lo que se conoce como una *función de activación* para computar su siguiente estado.(8) La manera más fácil de usar una red neuronal como una función es hacer que sea *de prealimentación*, de manera que la información fluya solo en una dirección, como en la figura 2.9, y se le proporcione a la capa superior la información de entrada, y la de salida se extraiga de la capa inferior de neuronas.

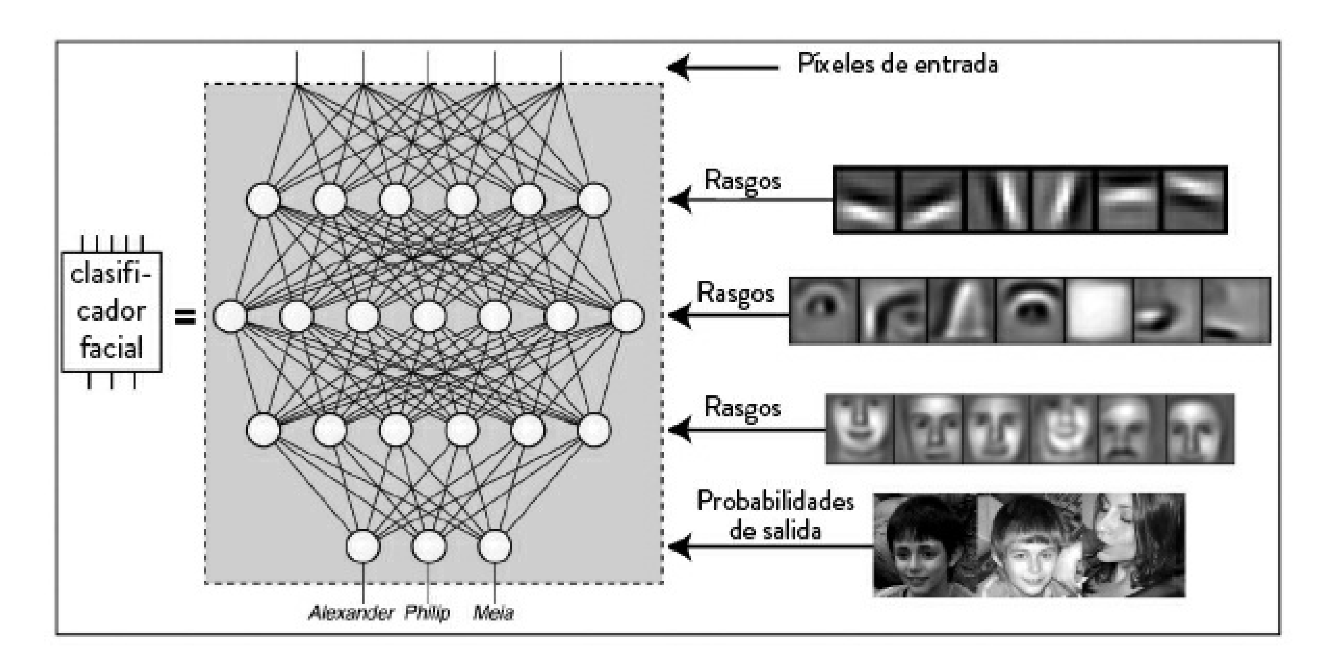


FIGURA 2.9. Una red de neuronas puede computar funciones igual que puede hacerlo una red de puertas NAND. Por ejemplo, algunas redes neuronales artificiales se han entrenado para tomar números que representan el brillo de distintos píxeles de una imagen y devolver números que representan la probabilidad de que la imagen retrate a una u otra persona. Aquí, cada neurona artificial (círculo) computa la suma ponderada de los números que le llegan a través de las conexiones (líneas) desde arriba, aplica una función simple y pasa el resultado hacia abajo, de forma que cada capa posterior computa rasgos de más alto nivel. Las redes de reconocimiento facial típicas contienen cientos de miles de neuronas; por claridad, en la figura solo aparecen unas cuantas.

El éxito de estas redes neuronales artificiales simples es un ejemplo más de la independencia respecto del sustrato: las redes neuronales tienen una gran potencia computacional aparentemente independiente de los detalles precisos de bajo nivel de cómo están construidas. De hecho, George Cybenko, Kurt Hornik, Maxwell Stinchcombe y Halbert White demostraron algo extraordinario en 1989: esas redes neuronales artificiales simples son *universales* en el sentido de que pueden computar *cualquier* función con la precisión que se quiera, ajustando debidamente los valores de las intensidades sinápticas. En otras palabras, es probable que la evolución no hiciera que nuestras neuronas biológicas fuesen tan complejas porque fuese necesario, sino porque era más eficiente; y porque la evolución, a diferencia de los ingenieros humanos, no prima los diseños sencillos y fáciles de entender.

Cuando me enteré de esto, me desconcertó saber que algo tan simple podía computar algo tan arbitrariamente complejo. Por ejemplo, ¿cómo podemos computar ni siquiera algo tan simple como una multiplicación cuando lo

único que se nos permite hacer es computar sumas ponderadas y aplicar una única función fija? Si quiere hacerse una idea de cómo funciona esto, la figura 2.10 muestra cómo tan solo cinco neuronas pueden multiplicar entre sí dos números arbitrariamente grandes, y cómo una sola neurona puede multiplicar entre sí tres bits.

Aunque podamos demostrar que es posible computar cualquier cosa en *teoría* con una red neuronal arbitrariamente grande, la demostración no dice nada sobre si podemos o no hacerlo en la *práctica* con una red de tamaño razonable. De hecho, cuanto más reflexionaba sobre ello, más me desconcertaba que las redes neuronales funcionasen tan bien.

Supongamos que queremos clasificar imágenes de un megapíxel en escala de grises en dos categorías; por ejemplo, gatos y perros. Si cada uno del millón de píxeles puede tomar uno de entre, pongamos, 256 valores, entonces hay $256^{1.000.000}$ imágenes posibles, y para cada una de ellas queremos computar la probabilidad de que represente un gato. Esto significa que una función arbitraria que recibe una imagen y devuelve una probabilidad viene definida por una lista de $256^{1.000.000}$ probabilidades, esto es, muchos más números que átomos hay en el universo (alrededor de 10^{78}). Pero redes neuronales con apenas miles o millones de parámetros de alguna manera logran realizar bastante bien tales tareas de clasificación. ¿Cómo es posible que redes neuronales efectivas sean tan «baratas», en el sentido de que requieran tan pocos parámetros? Al fin y al cabo, podemos demostrar que una red neuronal lo bastante pequeña para caber dentro del universo fracasará de forma estrepitosa a la hora de aproximar casi todas las funciones, y solo realizaría con éxito una fracción ridículamente pequeña de todas las tareas computacionales que se le podrían asignar.

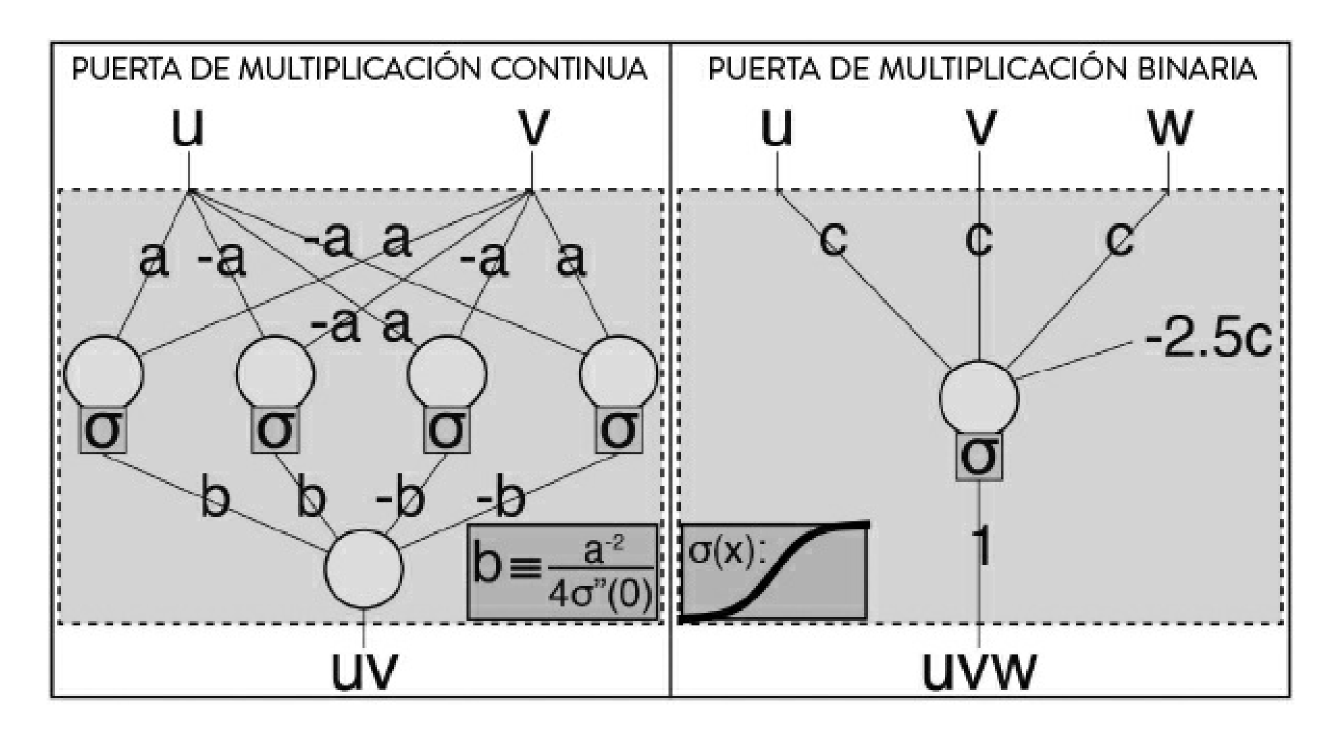


FIGURA 2.10. Cómo puede la materia multiplicar, pero no usando puertas NAND como en la figura 2.7, sino neuronas. Para entender la idea fundamental no es necesario seguir los detalles: se trata de que no solo las neuronas (artificiales o biológicas) pueden hacer matemáticas, sino que la multiplicación requiere muchas menos neuronas que puertas NAND. *Detalles opcionales para los muy interesados en las matemáticas*: Los círculos efectúan la suma, los cuadrados aplican la función σ, y las líneas multiplican por las constantes con que están etiquetadas. Las entradas son números reales (izquierda) y bits (derecha). La multiplicación se hace tan precisa como se quiera a medida que $a \longrightarrow 0$ (izquierda) y $c \longrightarrow \infty$ (derecha). La red de la izquierda funciona para cualquier función $\sigma(x)$ que sea curva en el origen (con segunda derivada $\sigma''(0) = 0$), que puede demostrarse desarrollando $\sigma(x)$ mediante una serie de Taylor. La red de la derecha requiere que la función $\sigma(x)$ se aproxime a 0 y a 1 cuando x es muy pequeña y muy grande, respectivamente, lo cual se ve si tenemos en cuenta que $uvw = 1$ solo si $u + v + w = 3$. (Estos ejemplos son de un artículo que escribí con mis estudiantes Henry Lin y David Rolnick, «Why Does Deep and Cheap Learning Work So Well?», que puede encontrarse en <http://arxiv.org/abs/1608.08225>.) Combinando un montón de multiplicaciones (como arriba) y adiciones, podemos computar cualquier polinomio, que, como es bien sabido, pueden aproximar cualquier función continuamente diferenciable.

Me he divertido muchísimo dando vueltas a este y otros misterios con mi estudiante Henry Lin. Una de las cosas por las que me siento más agradecido en la vida es la oportunidad de colaborar con estudiantes asombrosos, y Henry es uno de ellos. Cuando entró por primera vez en mi despacho para preguntarme si estaba interesado en trabajar con él, me dije que sería más apropiado que yo le preguntase si él tenía interés en trabajar conmigo: este chico modesto, simpático y de mirada brillante de Shreveport (Luisiana) ya había escrito ocho artículos científicos, ganado un premio 30-Under-30 de

Forbes y dado una charla TED con más de un millón de visualizaciones. ¡Y solo tenía veinte años! Un año más tarde, escribimos juntos un artículo con una conclusión sorprendente: la pregunta de por qué las redes neuronales funcionan tan bien no puede responderse tan solo usando matemáticas, porque parte de la respuesta está en la física. Descubrimos que la clase de funciones que las leyes de la física nos plantean y que hacen que nos interesemos por la computación es muy reducida porque, por razones que aún no comprendemos del todo, las leyes de la física son extraordinariamente simples. Además, la minúscula proporción de funciones que las redes neuronales pueden computar es muy parecida a la minúscula proporción en la que la física hace que nos interesemos. También ampliamos trabajos previos que demostraban que las redes neuronales que se usan en el aprendizaje profundo (se dice que son «profundas» si contienen más de una capa) son mucho más eficientes que las redes superficiales para muchas de estas funciones de interés. Por ejemplo, junto con otro fantástico estudiante del MIT, David Rolnick, demostramos que la sencilla tarea de multiplicar n números requiere nada menos que 2^n neuronas en una red con una sola capa, pero solo alrededor de $4n$ en una red profunda. Esto ayuda a explicar no solo por qué las redes neuronales están hoy en día tan en boga entre los investigadores en IA, sino también por qué desarrollamos redes neuronales en nuestro cerebro: si hemos desarrollado un cerebro para predecir el futuro, entonces tiene sentido que también hayamos desarrollado una arquitectura computacional que sea buena para abordar los problemas computacionales importantes en el mundo físico.

Ahora que hemos visto cómo funcionan y computan las redes neuronales, volvamos sobre la cuestión de cómo pueden aprender. Específicamente, ¿cómo puede una red neuronal mejorar su capacidad de computar actualizando sus sinapsis?

En *La organización de la conducta*, su libro seminal de 1949, el psicólogo canadiense Donald Hebb argumentó que, si dos neuronas próximas se activaban («se excitaban») con frecuencia al mismo tiempo, su acoplamiento sináptico se reforzaría de tal forma que aprenderían a excitarse mutuamente; el roce hace el cariño. Aunque aún estamos lejos de comprender los entresijos de cómo aprende un cerebro real, y las investigaciones han demostrado que las respuestas son en muchos casos mucho más complicadas, también se ha

demostrado que incluso esta sencilla regla de aprendizaje (conocida como aprendizaje hebbiano) hace posible que las redes neuronales aprendan cosas interesantes. John Hopfield demostró que el aprendizaje hebbiano permitía que su red neuronal artificial sobresimplificada almacenase una gran cantidad de recuerdos complejos con tan solo estar expuesta a ellos repetidamente. Dicha exposición a la información para aprender suele denominarse «entrenamiento» al hablar de redes neuronales artificiales (o de animales o personas a los que se les enseñan habilidades), aunque «estudio», «educación» o «experiencia» serían igualmente válidos. Las redes neuronales artificiales en las que se basan los sistemas actuales de IA suelen sustituir el aprendizaje hebbiano por otras reglas de aprendizaje más complejas, con nombres técnicos como «retropropagación» y «descenso del gradiente estocástico», pero la idea básica es la misma: existe alguna sencilla regla y determinista, análoga a una ley física, según la cual las sinapsis se actualizan a lo largo del tiempo. Como por arte de magia, esta sencilla regla puede hacer que la red neuronal aprenda computaciones en extremo complejas si se la entrena con grandes cantidades de datos. Aún no sabemos exactamente qué reglas utiliza nuestro cerebro pero, sean las que sean, no hay indicios de que violen las leyes de la física.

La mayoría de los ordenadores digitales mejoran su eficiencia al dividir su trabajo en varios pasos y reutilizar sus módulos computacionales muchas veces, y lo mismo hacen numerosas redes neuronales artificiales y biológicas. Los cerebros tienen partes que son lo que los informáticos llaman redes neuronales *recurrentes* en lugar de *prealimentadas*, en las cuales la información puede fluir en varias direcciones en lugar de hacerlo en una sola, de tal manera que la salida actual puede convertirse en entrada de lo que sucede a continuación. La red de puertas lógicas en el microprocesador de un ordenador portátil también es recurrente en este sentido: reutiliza continuamente su información anterior, y permite que nuevas entradas de información procedentes de un teclado, un *panel táctil*, una cámara, etcétera, afecte a la computación en curso, lo que a su vez determina la salida de información hacia, por ejemplo, un monitor, un altavoz, una impresora o una red inalámbrica. De forma análoga, la red de neuronas en nuestro cerebro es recurrente, y permite que la información procedente de los ojos, los oídos y otros sentidos afecte a la computación que está realizando, la cual a su vez determina la salida de información hacia nuestros músculos.

La historia del aprendizaje es al menos tan larga como la historia de la propia vida, ya que todo organismo autorreplicante lleva a cabo un interesante proceso de copiado y procesamiento de información, un comportamiento que ha aprendido de alguna manera. Durante la era de la vida 1.0, en cambio, los organismos no aprendían a lo largo de sus vidas: sus reglas para procesar la información y reaccionar venían determinadas por el ADN que habían heredado, por lo que el único aprendizaje que tenía lugar se producía lentamente y al nivel de la especie, a través de la evolución darwiniana a lo largo de generaciones.

Hace unos quinientos millones de años, ciertos linajes genéticos aquí en la Tierra descubrieron una manera de producir animales que contenían redes neuronales, capaces de aprender comportamientos a partir de las experiencias que acumulaban a lo largo de sus vidas. Había llegado la vida 2.0, y gracias a su capacidad de aprender de forma mucho más rápida y de ganarle la partida a la competencia, se extendió rápidamente por todo el mundo. Como vimos en el capítulo 1, la vida ha ido mejorando su capacidad de aprender, y a una velocidad cada vez mayor. Una especie concreta de simio desarrolló un cerebro tan competente para la adquisición de conocimiento que aprendió a usar herramientas, a hacer fuego, a hablar un lenguaje y a crear una sociedad compleja y global. A su vez, esta sociedad puede entenderse como un sistema que recuerda, computa y aprende, todo ello a un ritmo cada vez más acelerado a medida que una invención hace posible la siguiente: la escritura, la imprenta, la ciencia moderna, los ordenadores, internet, y así sucesivamente. ¿Qué será lo siguiente que pongan los historiadores del futuro en esa lista de inventos que abren posibilidades? Yo apuesto por la inteligencia artificial.

Como todos sabemos, las explosivas mejoras en cuanto a la memoria y capacidad de computación de los ordenadores (figuras 2.4 y 2.8) se han traducido en un avance espectacular en inteligencia artificial, pero tuvo que pasar mucho tiempo hasta que madurase el *aprendizaje* automático. Cuando el Deep Blue de IBM superó al campeón de ajedrez Garri Kaspárov en 1997, sus principales ventajas radicaban en la memoria y la capacidad de cómputo, no en el aprendizaje. Su inteligencia computacional había sido creada por un equipo de humanos, y la razón clave por la que Deep Blue pudo jugar mejor que sus creadores fue su capacidad para computar más rápido, y así poder analizar más posiciones potenciales. Cuando el ordenador Watson de IBM

destronó al campeón mundial humano en el concurso Jeopardy!, también se basó más en sus habilidades programadas específicamente para la ocasión y en una memoria y velocidad superiores que en el proceso de aprendizaje. Lo mismo puede decirse de la mayoría de los primeros avances en robótica, desde la locomoción usando prótesis hasta los coches autónomos y los cohetes capaces de aterrizar sin intervención humana.



FIGURA 2.11. «Un grupo de jóvenes jugando al plato volador.» Este pie de foto lo escribió un ordenador que no tiene ni idea de lo que son personas, juegos o platos voladores.

Por el contrario, la fuerza impulsora tras muchos de los avances más recientes en IA ha sido el *aprendizaje* automático. Fijémonos en la figura 2.11, por ejemplo. Para nosotros, es fácil decir de qué foto se trata, pero, durante décadas, ninguno de los investigadores en IA del mundo fue capaz de programar una función que no recibiese más que los colores de todos los píxeles de una imagen y devolviese un texto certero como «Un grupo de jóvenes jugando al plato volador». Pues eso es lo que hizo en 2014 un equipo de Google.[8] Si se introducen un conjunto distinto de colores de los píxeles, el programa responde de nuevo correctamente: «Una manada de elefantes atravesando una sabana seca». ¿Cómo lo consiguieron? ¿Al estilo de Deep

Blue, programando algoritmos artesanales para detectar platos voladores, rostros y demás? No. Lo hicieron creando una red neuronal bastante sencilla que carecía de cualquier conocimiento del mundo físico o de su contenido, y después dejando que esta aprendiese al exponerla a cantidades enormes de datos. El visionario de la IA Jeff Hawkins escribió en 2004 que «ningún ordenador puede […] ver tan bien como un ratón», pero hace ya mucho tiempo que eso no es así.

Así como no entendemos del todo cómo aprenden nuestros hijos, tampoco comprendemos plenamente aún cómo aprenden estas redes neuronales, y por qué fallan en ocasiones. Pero lo que está claro es que son muy útiles y están dando lugar a una oleada de inversiones en aprendizaje profundo. El aprendizaje profundo ha transformado muchos aspectos de la visión artificial, desde la transcripción de textos escritos a mano hasta el análisis de vídeos en tiempo real en los coches autónomos. También ha revolucionado la capacidad de los ordenadores para transformar el lenguaje hablado en texto y traducirlo a otros idiomas, incluso en tiempo real (por eso podemos hablar con asistentes personales digitales como Siri, Google Now y Cortana). Los molestos CAPTCHA mediante los que debemos convencer a un sitio web de que somos humanos son cada vez más difíciles para mantenerse por delante de lo que la tecnología de aprendizaje automático es capaz de hacer. En 2015, Google DeepMind lanzó un sistema de IA que utilizaba aprendizaje profundo y era capaz de dominar decenas de juegos de ordenador como lo haría un chaval —sin recibir ni una sola instrucción—, pero que enseguida aprendió a jugar mejor que cualquier humano. En 2016, la misma compañía construyó AlphaGo, un sistema informático que jugaba al go y usaba el aprendizaje profundo para evaluar la fortaleza de las distintas posiciones en el tablero y venció al mejor jugador del mundo. Estos avances están alimentando un circulo virtuoso que hace que la investigación en IA reciba cada vez más financiación y talento, lo que da lugar a más avances.

Hemos dedicado este capítulo a explorar la naturaleza de la inteligencia y su desarrollo hasta el momento actual. ¿Cuánto tiempo pasará hasta que las máquinas nos superen en todas las tareas cognitivas? Está claro que no lo sabemos, y debemos estar abiertos a la posibilidad de que la respuesta sea «nunca». Sin embargo, un mensaje básico de este capítulo es que también hemos de considerar la posibilidad de que suceda, quizá incluso a lo largo de nuestras vidas. A fin de cuentas, la materia se puede organizar de manera que,

al tiempo que cumple las leyes físicas, recuerde, compute y aprenda; y la materia no tiene por qué ser biológica. Con frecuencia, se acusa a los investigadores en IA de un exceso de optimismo en sus promesas y de una escasez de resultados reales, pero, si hemos de ser justos, algunos de sus detractores tampoco tienen un historial intachable. Hay quien no deja de cambiar las reglas del juego en mitad de la partida, y definen la inteligencia como aquello que los ordenadores aún no son capaces de hacer, o como lo que nos impresiona. Las máquinas ahora son buenas o excelentes en aritmética, ajedrez, demostración de teoremas matemáticos, selección de valores bursátiles, elaboración de pies de fotos, conducción, videojuegos, go, síntesis de voz, transcripción del habla, traducción y diagnóstico del cáncer; sin embargo, algunos críticos se mofarán desdeñosamente diciendo: «Claro, ¡pero eso no es verdadera inteligencia!». Luego añadirán que la única inteligencia real es la que tiene que ver con las cumbres del paisaje de Moravec (figura 2.2) que aún no están bajo el agua, como hubo gente en el pasado que argumentó que los pies de foto y el go deberían incluirse… mientras el agua seguía creciendo.

Suponiendo que el agua seguirá subiendo al menos durante un tiempo, el impacto de la IA en la sociedad no dejará de crecer. Mucho tiempo antes de que la IA alcance un nivel humano en todas las tareas, nos ofrecerá fascinantes oportunidades y desafíos relacionados con cuestiones como depuración de programas, leyes, armas y puestos de trabajo. ¿Cuáles son esas oportunidades y desafíos y cuál es la mejor manera de prepararnos para ellos? Eso es lo que veremos en el capítulo siguiente.

CONCLUSIONES

- La inteligencia, definida como la capacidad de lograr objetivos complejos, no puede medirse por medio de un solo CI, sino a través de un espectro de capacidades respecto a todos los objetivos.
- La inteligencia artificial actual suele ser *estrecha*: cada sistema solo es capaz de lograr objetivos muy específicos, mientras que la inteligencia humana es extraordinariamente *amplia*.
- Memoria, computación, aprendizaje e inteligencia tienen todos un aire abstracto, intangible y etéreo porque son *independientes del sustrato*: pueden tomar vida propia que no depende del sustrato material subyacente ni refleja los detalles del mismo.
- Siempre que tenga muchos estados estables distintos, cualquier pedazo de materia puede ser el sustrato de una *memoria*.
- Cualquier materia puede ser *computronio*, el sustrato para la *computación*, siempre que contenga

determinados elementos universales que puedan combinarse para implementar cualquier función. Las puertas NAND y las neuronas son dos ejemplos importantes de estos «átomos computacionales» universales.

- Una red neuronal es un sustrato potente para el *aprendizaje* porque, por el hecho de obedecer las leyes de la física, puede reorganizarse para mejorar progresivamente su capacidad de implementar determinadas funciones.
- Debido a la llamativa simplicidad de las leyes de la física, a los humanos solo nos interesa una proporción minúscula de todos los problemas computacionales imaginables, y las redes neuronales suelen ser en extremo eficaces para resolver esta minúscula proporción.
- Una vez que la tecnología es el doble de potente, a menudo puede utilizarse para diseñar y construir tecnología que es a su vez el doble de potente, lo que desencadena una duplicación reiterada de la capacidad, en la línea de la ley de Moore. El coste de la tecnología de la información se ha reducido a la mitad aproximadamente cada dos años durante un siglo, haciendo posible la era de la información.
- Si continúa el avance de la IA, entonces, mucho tiempo antes de que la IA alcance un nivel humano en todas sus habilidades, nos proporcionará fascinantes oportunidades y desafíos relacionados con cuestiones como depuración de programas, leyes, armas y puestos de trabajo, cosa que abordaremos en el capítulo siguiente.

3
EL FUTURO PRÓXIMO
AVANCES, GAZAPOS, LEYES, ARMAS Y PUESTOS DE TRABAJO

Si no enderezamos el rumbo pronto, acabaremos allí donde estamos yendo.
IRWIN COREY

¿Qué significa ser humano en la actualidad? Por ejemplo, ¿qué es lo que realmente valoramos de nosotros mismos, lo que nos hace diferentes de otras formas de vida y de las máquinas? ¿Qué es lo que otras personas valoran de nosotros para que algunas de ellas estén dispuestas a ofrecernos un trabajo? Sean cuales sean nuestras respuestas a estas preguntas en un momento dado, está claro que la irrupción de la tecnología debe ir cambiándolas poco a poco.

Consideremos mi propio caso. Como científico, me enorgullezco de marcarme mis propios objetivos, de aplicar creatividad e intuición para abordar un amplio abanico de problemas aún por resolver, y en emplear el lenguaje para difundir lo que descubro. Por fortuna para mí, la sociedad está dispuesta a pagarme por hacer este trabajo. Hace siglos, como muchos otros, habría sido campesino o artesano, pero el crecimiento de la tecnología desde entonces ha hecho que esas profesiones constituyan ahora una minúscula proporción de la mano de obra. Esto significa que ya no es posible que todo el mundo se dedique a la agricultura o a la artesanía.

Personalmente, no me molesta que las máquinas actuales me superen en habilidades manuales como cavar o tejer, puesto que estas no son pasatiempos que practique ni mis fuentes de ingresos o de autoestima. De hecho, cualquier falsa ilusión que me hubiese hecho sobre mis habilidades a ese respecto fue aplastada a la edad de ocho años, cuando mi colegio me obligó a dar una asignatura de hacer punto que estuve a punto de suspender, y conseguí completar mi proyecto gracias a la ayuda de un compañero compasivo de quinto curso que se apiadó de mí.

Pero, a medida que la tecnología sigue mejorando, ¿el ascenso de la IA acabará por hacer sombra a todas esas capacidades en las que se basa hoy en

día mi autoestima y me otorgan valor en el mercado laboral? Stuart Russell me contó que él y muchos de sus colegas investigadores en IA habían experimentado recientemente la necesidad de soltar un «¡hostia!» al presenciar cómo la IA hacía algo que no esperaban ver aún en muchos años. En ese sentido, permítame que le hable de algunos momentos similares que yo he experimentado y por qué los considero presagios de que ciertas capacidades humanas serán pronto superadas.

AVANCES

Agentes de aprendizaje profundo por refuerzo

Uno de mis momentos de mayor asombro lo viví en 2014 mientras veía un vídeo en el que un sistema de IA de DeepMind aprendía a jugar a videojuegos. En particular, la IA estaba jugando a *Breakout* (véase la figura 3.1), un juego clásico de Atari que recuerdo con cariño de mi adolescencia. El objetivo es manejar una pala para hacer que una pelota rebote repetidamente contra un muro de ladrillos; cada vez que se golpea un ladrillo, este desaparece y aumenta la puntuación.

Había programado juegos de ordenador por mi cuenta tiempo atrás, y era consciente de que no resultaba difícil escribir un programa capaz de jugar a *Breakout*, pero no era eso lo que el equipo de DeepMind había hecho: ellos habían creado una IA *tabula rasa* que no sabía nada sobre este juego o sobre ningún otro, ni siquiera sobre *conceptos* como los de juego, pala, ladrillo o pelota. Lo único que sabía su IA era que recibía una larga lista de números a intervalos regulares: la puntuación actual y una larga lista de números que nosotros reconoceríamos (pero la IA no) como especificaciones de cómo estaban coloreadas distintas partes de la pantalla. A la IA se le dijo que maximizase la puntuación devolviendo, a intervalos regulares, números que nosotros (pero no la IA) reconoceríamos como códigos de qué teclas habría que pulsar.

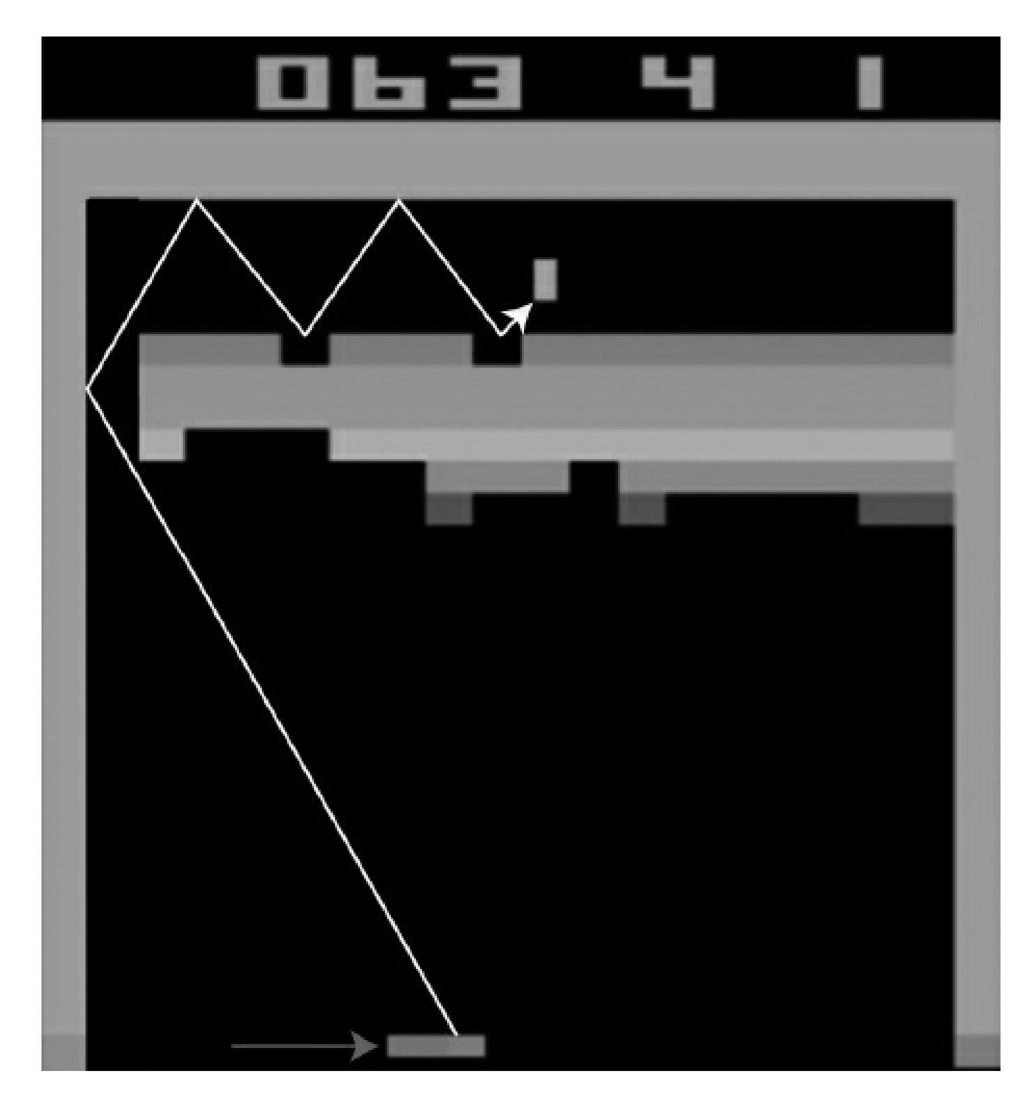


FIGURA 3.1. Tras aprender a jugar al juego de Atari desde cero, usando aprendizaje profundo por refuerzo para maximizar la puntuación, la IA de DeepMind descubrió la estrategia óptima: hacer un agujero en la esquina superior izquierda del muro de ladrillos y dejar que la pelota se ponga a rebotar tras él, lo que permite acumular puntos muy rápidamente. Las flechas que he dibujado representan las trayectorias pasadas de la pelota y de la pala.

Al principio, la IA jugó fatal: movía la pala de un lado a otro de forma aparentemente aleatoria y sin ningún criterio y casi nunca conseguía golpear la pelota. Al cabo de un rato, parecía que iba entendiendo que era buena idea mover la pala hacia la pelota, aunque aún fallaba la mayoría de las veces. Pero siguió mejorando con la práctica, y enseguida llegó a jugar mejor de lo que yo nunca lo hice, y siempre devolvía la bola por muy rápido que esta se moviese. Entonces ocurrió algo inaudito: dedujo una asombrosa forma de maximizar la puntuación consistente en apuntar siempre hacia la esquina

superior izquierda para hacer un agujero que atravesara el muro, de manera que la pelota se quedase atrapada rebotando entre la parte posterior del muro y la barrera existente tras él. De hecho, Demis Hassabis me dijo más tarde que los programadores que formaban parte de ese equipo de DeepMind no conocían este truco hasta que lo aprendieron de la IA que habían construido. Le recomiendo que vea por usted mismo un vídeo del proceso en el enlace que incluyo aquí.[9]

Había en la situación algo de humano que me resultaba un tanto desazonante: estaba viendo una IA que tenía un objetivo y aprendía a mejorar en su capacidad de lograrlo, hasta acabar haciéndolo mejor que sus creadores. En el capítulo anterior, definimos la inteligencia como la capacidad de lograr objetivos complejos, por lo que, en este sentido, la IA de DeepMind se estaba volviendo cada vez más inteligente ante mis ojos (aunque únicamente en su capacidad para jugar a este juego en particular). En el primer capítulo vimos lo que los informáticos llaman *agentes inteligentes*: entidades que captan información de su entorno a través de sensores y a continuación la procesan para decidir cómo actuar sobre dicho entorno. Aunque la IA jugadora de DeepMind vivía en un mundo muy simple compuesto de ladrillos, palas y pelotas, no podía negar que se trataba de un agente inteligente.

DeepMind enseguida publicó su método y compartió su código, explicando que usaba una idea muy sencilla, pero potente, llamada *aprendizaje profundo por refuerzo*.[10] El aprendizaje por refuerzo básico es una técnica clásica de aprendizaje automático inspirada en la psicología conductista, según la cual la obtención de una recompensa positiva aumenta la probabilidad de que uno haga algo de nuevo, y viceversa. De la misma manera en que un perro aprende a hacer trucos cuando esto incrementa la probabilidad de recibir poco después algún gesto de aliento o una golosina por parte de su amo, la IA de DeepMind aprendió a mover la pala para devolver la pelota porque esto incrementaba la probabilidad de obtener una mayor puntuación. DeepMind combinó esta idea con el aprendizaje profundo: el equipo entrenó una red neuronal profunda, como las del capítulo anterior, para que predijese cuántos puntos se obtendrían de media al presionar cada una de las teclas permitidas del teclado, y a continuación la IA seleccionaba aquella tecla que la red neuronal había calificado como más prometedora dado el estado del juego en cada momento.

Cuando enumeré los rasgos que contribuyen a mi propia sensación de

autoestima como humano, incluí la capacidad de abordar una amplia variedad de problemas aún por resolver. Por el contrario, ser capaz de jugar tan solo a *Breakout* constituye una inteligencia sumamente estrecha. Para mí, la verdadera importancia del avance de DeepMind es que el aprendizaje profundo por refuerzo es una técnica de aplicación completamente general. Como era de esperar, hicieron que esa misma IA intentase aprender a jugar a cuarenta y nueve juegos de Atari distintos, y superó a los evaluadores humanos en veintinueve de ellos, desde *Pong* hasta *Boxing, Video Pinball* y *Space Invaders.*

No pasó mucho tiempo antes de que esa misma idea empezase a probarse en juegos más modernos, cuyos mundos eran tridimensionales en lugar de bidimensionales. Enseguida, OpenAI, unos competidores de DeepMind con sede en San Francisco, lanzaron una plataforma llamada Universe, en la que la IA de DeepMind y otros agentes inteligentes pueden practicar a interactuar con un ordenador entero como si fuese un juego: pulsando cualquier cosa, escribiendo lo que sea y abriendo y ejecutando cualquier software en el que sepan moverse (por ejemplo, arrancando un navegador web y perdiendo el tiempo online).

Al mirar hacia el futuro del aprendizaje profundo por refuerzo y las mejoras que experimentará, no parece que tengan ningún límite evidente. Su potencial no se limita a los mundos virtuales de los videojuegos, puesto que, si uno es un robot, puede ver la propia vida como un juego. Stuart Russell me contó que tuvo su primer momento ¡hostia! cuando vio cómo el robot Big Dog subía una cuesta cubierta de nieve en medio de un bosque, resolviendo así elegantemente el problema de la locomoción con extremidades que él mismo había tratado de solventar durante muchos años.[11] Pero, cuando ese hito se alcanzó en 2008, requirió una enorme cantidad de trabajo por parte de ingeniosos programadores. Tras el avance de DeepMind, no hay motivo para que un robot no use alguna variante del aprendizaje profundo por refuerzo para aprender por su cuenta a andar sin ayuda de programadores humanos: todo lo que se necesita es un sistema que le dé puntos cada vez que haga un progreso. Los robots en el mundo real son también capaces de aprender a nadar, volar, jugar al ping-pong, combatir y realizar una tarea casi interminable de otras tareas motrices sin ayuda de programadores humanos. Para acelerar las cosas y reducir el riesgo de que se atasquen o sufran daños durante el proceso de aprendizaje, probablemente llevarían a cabo las

primeras fases de su aprendizaje en un entorno de realidad virtual.

Intuición, creatividad y estrategia

Otro momento determinante para mí fue cuando AlphaGo, un sistema de IA de DeepMind, ganó un enfrentamiento de go a cinco partidas contra Lee Sedol, considerado como el mejor jugador del mundo a principios del siglo XXI.

Mucha gente esperaba que los jugadores de go humanos fuesen destronados por las máquinas en algún momento, ya que eso mismo había sucedido con los jugadores de ajedrez dos décadas antes. Sin embargo, la mayoría de los expertos en go predecían que eso tardaría aún en torno a una década en suceder, por lo que el triunfo de AlphaGo fue un momento tan decisivo para ellos como lo fue para mí. Tanto Nick Bostrom como Ray Kurzweil han insistido en lo difícil que es prever estos avances en IA, lo cual resulta evidente de las entrevistas con el propio Lee Sedol antes y después de perder las tres primeras partidas:

- Octubre de 2015: «Viendo el nivel que ha demostrado […] creo que ganaré la partida casi de forma arrolladora».
- Febrero de 2016: «He oído que la IA de Google DeepMind es sorprendentemente potente, y cada vez más, pero confío en poder vencerla al menos esta vez».
- 9 de marzo de 2016: «Me quedé muy sorprendido, porque no esperaba perder».
- 10 de marzo de 2016: «No tengo palabras […] Estoy en shock. Reconozco […] que la tercera partida no va a ser fácil para mí».
- 12 de marzo de 2016: «Sentí una especie de impotencia».

Un año después de enfrentarse a Lee Sedol, una versión mejorada de AlphaGo había jugado contra los veinte mejores jugadores del mundo sin perder una sola partida.

¿Por qué fue esto tan importante para mí? Antes confesé que considero que la intuición y la creatividad son dos de mis características fundamentales como humano, y, como explicaré a continuación, sentí que AlphaGo había hecho muestra de ambas.

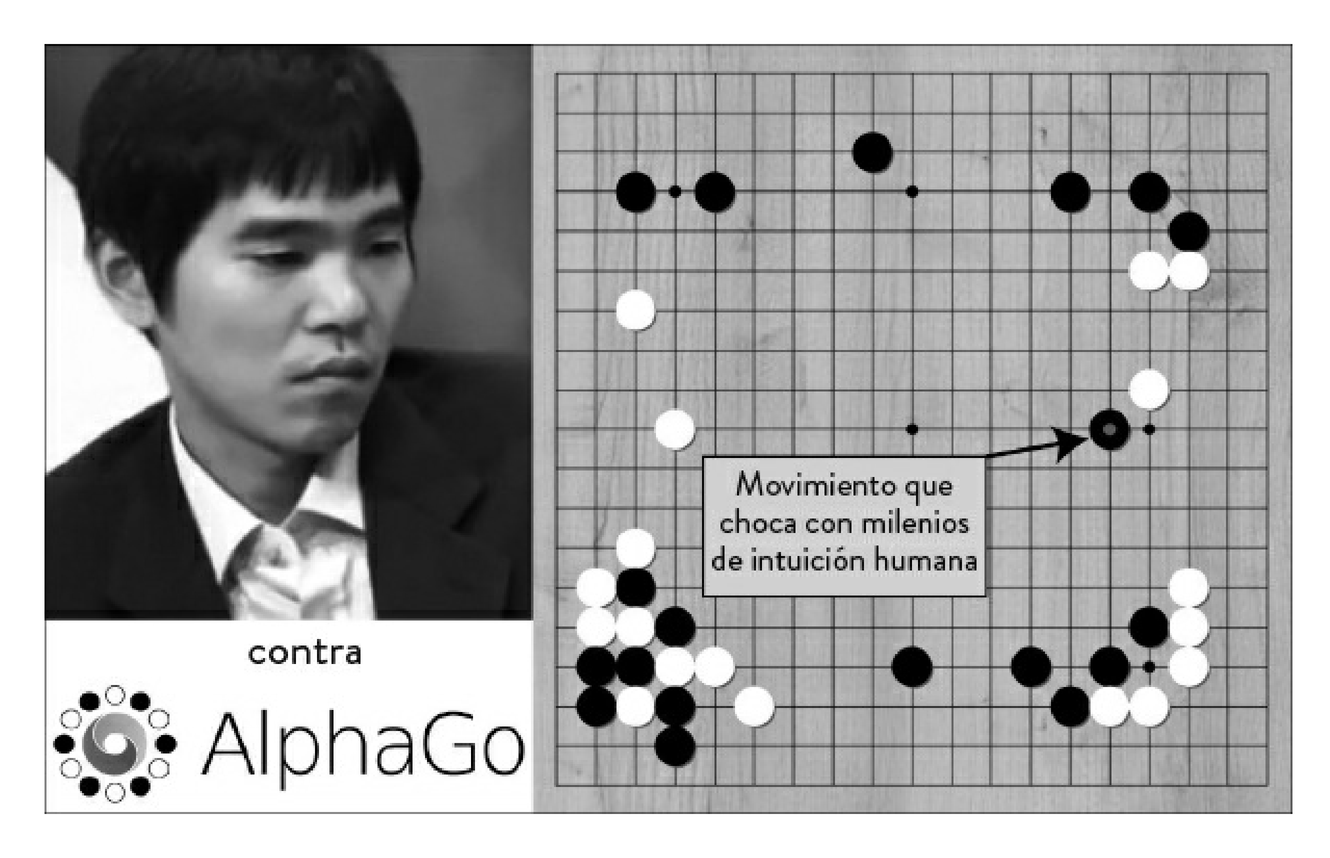


FIGURA 3.2. La IA AlphaGo de DeepMind hizo un movimiento sumamente creativo en la línea 5, que iba contra milenios de sabiduría humana y que cincuenta movimientos después resultó ser crucial para su victoria sobre la leyenda del go Lee Sedol.

Los jugadores de go se turnan para colocar piedras negras y blancas en un tablero de 19 por 19 casillas (véase la figura 3.2). La cantidad de posiciones posibles en go es muchísimo mayor que la de átomos en el universo, lo que significa que intentar analizar todas las secuencias interesantes de movimientos futuros es una tarea abocada al fracaso. Por ese motivo, los jugadores se basan en gran medida en su intuición para complementar sus razonamientos conscientes, y los expertos desarrollan un talento casi sobrenatural para saber qué posiciones son fuertes y cuáles son débiles. Como vimos en el capítulo anterior, los resultados del aprendizaje profundo a veces dan la sensación de ser fruto de la intuición: una red neuronal profunda podría determinar que una imagen representa un gato sin ser capaz de explicar por qué. Así pues, el equipo de DeepMind apostó por la idea de que el aprendizaje profundo podría ser capaz de reconocer no solo gatos sino también posiciones ventajosas de go. La idea central que integraron en AlphaGo fue la de combinar la potencia intuitiva del aprendizaje profundo

con la capacidad lógica de la GOFAI (siglas de lo que se conoce jocosamente como «Good Old-Fashioned IA», la «IA de toda la vida», anterior a la revolución que ha supuesto el aprendizaje profundo). Usaron una enorme base de datos de posiciones de go procedentes tanto de partidas jugadas por humanos como de otras en las que AlphaGo había jugado contra un clon de sí mismo, y entrenaron a una red neuronal profunda para predecir, a partir de cada posición, la probabilidad de que las fichas blancas acabasen ganando. También entrenaron a otra red distinta para que predijese qué jugadas eran más probables a continuación de una dada. Después, combinaron ambas redes mediante un método de GOFAI que buscaba ingeniosamente en una lista filtrada de secuencias probables de movimientos futuros para identificar el movimiento siguiente que conduciría a la posición más fuerte más adelante.

Esta combinación de intuición y lógica tuvo como consecuencia movimientos que no solo eran potentes sino, en algunos casos, sumamente creativos. Por ejemplo, milenios de sabiduría acumulada en torno al go dictan que, en las primeras jugadas de una partida, es preferible jugar en la tercera o cuarta línea contadas desde un extremo. Cada opción tiene sus ventajas: jugar en la tercera línea ayuda a ganar territorio a corto plazo hacia el borde, mientras que hacerlo en la cuarta contribuye a la influencia estratégica a largo plazo hacia el centro.

En el trigésimo séptimo movimiento de la segunda partida, AlphaGo asombró al mundo del go al hacer caso omiso de esa sabiduría ancestral y jugar en la quinta línea (figura 3.2), como si tuviera más confianza aún que un humano en su capacidad de planificación a largo plazo, y por tanto prefiriese optar por la ventaja estratégica frente a la ganancia a corto plazo. Los comentaristas se quedaron perplejos, y Lee Sedol llegó incluso a levantarse y salir de la sala.[12] Como cabía esperar, unos cincuenta movimientos más tarde, la lucha en la esquina inferior izquierda del tablero acabó extendiéndose y conectando con esa piedra negra del movimiento 37. Y ese detalle fue el que acabó decidiendo la partida, y haciendo que el movimiento de AlphaGo en la quinta fila pasase a la posteridad como uno de los más creativos de la historia del go.

Debido a sus facetas intuitiva y creativa, el go suele verse más como un arte que como otro juego cualquiera. En la antigua China era considerado uno de los cuatro «artes esenciales», junto con la pintura, la caligrafía y la música

qin, y aún hoy en día sigue siendo enormemente popular en Asia, donde casi trescientos millones de personas vieron la primera partida entre AlphaGo y Lee Sedol. En consecuencia, el resultado del enfrentamiento provocó una sacudida en el mundo del go, que interpretó la victoria de AlphaGo como un importante hito para la humanidad. Ke Jie, que por aquel entonces lideraba la clasificación mundial de go, dijo al respecto:[13] «La humanidad lleva miles de años jugando al go y pesar de ello, como la IA nos ha demostrado, aún no hemos hecho más que empezar a entenderlo […]. La unión de jugadores humanos y ordenadores supondrá el comienzo de una nueva era […]. Juntos, el hombre y la IA podrán descubrir la verdad del go». Esa fructífera colaboración entre humanos y máquinas parece efectivamente prometedora en muchas áreas, incluida la ciencia, donde la IA podría ayudarnos a los humanos a alcanzar una comprensión más profunda y hacer realidad todo nuestro potencial.

En 2017 el equipo de DeepMind lanzó el sucesor de AlphaGo, AlphaZero. Tomó los miles de años de sabiduría humana en go, incluyendo millones de partidas, y los ignoró por completo; aprendió desde cero simplemente jugando consigo mismo. No solo ganó a AlphaGo, sino que aprendió hasta convertirse en el mejor jugador de ajedrez del mundo solo jugando consigo mismo. Después de dos horas de práctica pudo vencer a los mejores jugadores humanos, y tras cuatro horas destronó al mejor programa de ajedrez. Lo que me pareció más sorprendente fue que, además de vencer a jugadores humanos de ajedrez, consiguió hacer lo mismo con los programadores de IA, dejando obsoleto todo el trabajo manual en IA que habían desarrollado durante décadas. En otras palabras, no podemos desechar la idea de IA creando mejores IA.

En mi opinión, AlphaGo nos enseña también otra importante lección para el futuro próximo: la combinación de la intuición propia del aprendizaje profundo con la lógica de la GOFAI puede dar como resultado una *estrategia* inmejorable. Puesto que el go es uno de los más refinados juegos de estrategia, la IA está ahora en disposición de enfrentarse (o ayudar) a los mejores estrategas humanos más allá de los juegos de mesa; por ejemplo, a la hora de elaborar estrategias de inversión, políticas o militares. Estos problemas relacionados con las estrategias en el mundo real suelen verse complicados por la psicología humana, la carencia de información y factores que deben modelarse como aleatorios, pero los sistemas de IA para jugar al

póker ya han demostrado que ninguna de estas dificultades es insuperable.

Lenguaje natural

Otra área más en la que el progreso de la IA me ha sorprendido recientemente es el lenguaje. Desde muy joven he sido un enamorado de los viajes, y la curiosidad por otras culturas e idiomas ha sido una parte importante de mi identidad. Me eduqué hablando sueco e inglés, me enseñaron alemán y español en la escuela, aprendí portugués y rumano gracias a sendos matrimonios y estudié por mi cuenta algo de ruso, francés y mandarín solo por divertirme.

Pero la IA ha estado alcanzando, y después de un descubrimiento importante en 2016, casi no hay lenguajes perezosos que puedo traducir entre mejor que el sistema de inteligencia artificial desarrollado por el equipo del cerebro de Google.

¿Me he expresado con claridad? Lo que estaba tratando de decir era esto:

Pero la IA ha ido ganándome terreno y, tras un avance muy importante en 2016, ya casi no quedan idiomas entre los que yo pueda traducir mejor que el sistema de IA desarrollado por el equipo de Google Brain.

Sin embargo, primero lo traduje al español y de vuelta al inglés usando una aplicación que instalé hace unos años en mi portátil. En 2016, el equipo de Google Brain introdujo mejoras en su servicio gratuito Google Service y empezó a utilizar redes neuronales profundas recurrentes, y la mejoría respecto a los anteriores sistemas de GOFAI fue espectacular:[14]

Pero la IA ha ido ganándome terreno, y después de un avance en 2016, ya casi no quedan idiomas entre los que se pueda traducir mejor que el Sistema de IA desarrollado por el equipo de Google Brain.

Como podemos ver, el pronombre «yo» se perdió en el rodeo a través del español, lo que lamentablemente altera el significado del texto. ¡Casi bien, pero no del todo! No obstante, debo decir en descarga de la IA de Google que suelen criticarme porque escribo frases innecesariamente largas que dificultan su comprensión, y elegí una de las más enrevesadas para este ejemplo. Si las frases son más sencillas, su IA suele traducirlas de forma impecable. En consecuencia, creó cierto revuelo cuando se lanzó, y es lo bastante útil para que cientos de millones de personas lo usen cada día. Más aún, gracias a los

recientes avances en aprendizaje profundo para la conversión entre habla y texto y viceversa, hoy en día esos usuarios pueden hablarles a sus teléfonos inteligentes en un idioma y escuchar la traducción.

El procesamiento del lenguaje natural es ahora uno de los campos de la IA en el que se están produciendo avances más rápidos, y creo que sus éxitos futuros tendrán un gran impacto, porque el lenguaje es algo esencial para el ser humano. Cuanto mejor sea la capacidad de predicción lingüística de una IA, mejor podrá redactar respuestas razonables por correo electrónico o mantener una conversación oral. Esto podría dar la impresión, al menos a una persona que desconozca este campo, de que se está produciendo pensamiento humano. Así que los sistemas de aprendizaje profundo se aproximan poco a poco a posibilidad de superar el famoso test de Turing, según el cual una máquina debe conversar suficientemente bien por escrito con una persona para convencerla de que está comunicándose con un humano.

Pero la IA para el procesamiento del lenguaje aún tiene un largo camino por recorrer. Aunque debo confesar que me deprime un poco que una IA sea capaz de traducir mejor que yo, me siento mejor cuando me digo que, hasta ahora, no entiende lo que está diciendo en ningún sentido significativo. A base de entrenarse con infinidad de datos, descubre patrones y relaciones entre las palabras, sin establecer en ningún momento vínculos entre dichas palabras del mundo real. Por ejemplo, podría representar cada palabra mediante una lista de mil números que especifiquen su grado de semejanza con determinadas palabras, y de ello podría entonces deducir que la diferencia entre «rey» y «reina» es similar a la que existe entre «marido» y «mujer», pero sigue sin tener ni idea de lo que significa ser macho o hembra, o ni siquiera de que existe algo como una realidad física con espacio, tiempo y materia.

Puesto que el test de Turing gira básicamente en torno a la capacidad de engañar, hay quien lo critica porque lo que pone a prueba es la credulidad humana más que si hay una verdadera inteligencia artificial. Por el contrario, un test rival conocido como *desafío de los esquemas de Winograd* va directo a la yugular, y se centra en la comprensión asociada al sentido común de la que suelen carecer los actuales sistemas de aprendizaje profundo. Normalmente, los humanos usamos nuestro conocimiento del mundo real a la hora de interpretar una frase para determinar a qué hace referencia un pronombre o una persona de un verbo. Por ejemplo, un desafío de Winograd

típico pregunta a quién hace referencia el segundo verbo aquí:

1. Los concejales denegaron un permiso a los manifestantes porque temían la violencia.
2. Los concejales denegaron un permiso a los manifestantes porque defendían la violencia.

Existe una competición anual para responder a cuestiones como esta, y las IA obtienen aún pésimos resultados.[15] Este desafío en concreto, entender a qué se refiere cada cosa, hizo que incluso Google Translate fracasase cuando sustituí el español por el chino en mi ejemplo anterior:

Pero la inteligencia artificial me ha alcanzado, después de un gran descanso en 2016, con casi ningún idioma, pude traducir el sistema de inteligencia artificial desarrollado por el equipo de Google Brain.

Por favor, haga usted la prueba en <https://translate.google.com> ahora que está leyendo el libro y compruebe si la IA de Google ha mejorado. Es muy probable que así sea, ya que se han propuesto estrategias prometedoras para combinar las redes neuronales profundas recurrentes con la GOFAI para construir una IA orientada al procesamiento del lenguaje que incorpore un modelo del mundo.

Oportunidades y dificultades

Obviamente, estos no son más que tres ejemplos entre muchos, ya que la IA está avanzando más rápido en muchos frentes importantes. Además, aunque en los ejemplos solo he mencionado dos compañías, a menudo la competencia de grupos de investigación universitarios y de otras empresas no les iba muy a la zaga. En los departamentos de informática de universidades de todo el mundo puede notarse el efecto de succión que ejercen Apple, Baidu, DeepMind, Facebook, Google, Microsoft y otras empresas que ofrecen lucrativos contratos para atraer a estudiantes, doctores y profesores.

Es importante que los ejemplos que he puesto no nos lleven a engaño y veamos la historia de la IA como una sucesión de periodos de estancamiento salpicados por algún que otro avance importante. Por el contrario, lo que yo he observado es un progreso bastante sostenido durante mucho tiempo que los medios de comunicación presentan como un gran avance cada vez que da lugar a una nueva aplicación o a un producto útil capaces de llamar nuestra

atención. Habida cuenta de lo cual considero probable que el progreso de la IA continúe a este ritmo vivo durante muchos años. Además, como vimos en el capítulo anterior, no hay ninguna razón fundamental para que este progreso no pueda continuar hasta que la IA iguale la capacidad humana en la mayoría de tareas.

Lo cual suscita una pregunta: ¿cómo nos afectará esto? ¿Cómo alterará el progreso a corto plazo de la IA lo que significa ser humano? Hemos visto que cada vez es más difícil argumentar que la IA carece por completo de objetivos, amplitud, intuición, creatividad o lenguaje, rasgos que muchos consideran consustanciales al ser humano. Esto implica que, incluso a corto plazo, mucho antes de que cualquier IAG pueda alcanzar nuestro nivel en todas las tareas, la IA podría tener un impacto considerable sobre cómo nos vemos a nosotros mismos, lo que podemos llevar a cabo en cooperación con la IA y sobre qué podemos hacer para ganar dinero cuando tengamos que competir contra la IA. ¿Ese impacto será positivo o negativo? ¿Qué oportunidades y dificultades presentará esta situación?

Todo lo que nos gusta de la civilización es producto de la inteligencia humana, por lo que, si podemos amplificarla mediante la inteligencia artificial, obviamente se abrirá ante nosotros la posibilidad de hacer que la vida sea aún mejor. Incluso un modesto avance en la IA podría traducirse en mejoras considerables en ciencia y tecnología, con la consiguiente reducción de los accidentes, las enfermedades, la injusticia, la guerra, el trabajo pesado y la pobreza. Pero, para poder cosechar estos beneficios de la IA sin crear nuevos problemas, necesitamos dar respuesta a muchas preguntas importantes. Por ejemplo:

1. ¿Cómo podemos hacer que los futuros sistemas de IA sean más robustos que los actuales, para que hagan lo que queremos sin colapsar, tener fallos o ser pirateados?
2. ¿Cómo podemos modificar nuestros sistemas legales para que sean más justos y eficientes y no pierdan comba respecto a los cambios tan rápidos que se producen en el panorama digital?
3. ¿Cómo podemos crear armas más inteligentes y menos propensas a matar a civiles inocentes sin desencadenar una descontrolada carrera armamentística de armas autónomas letales?
4. ¿Cómo podemos acrecentar nuestra prosperidad a través de la automatización sin privar a la gente de ingresos o del sentido de su vida?

Dedicaremos el resto del capítulo a abordar cada una de estas cuestiones. Estas cuatro preguntas a corto plazo van dirigidas sobre todo a informáticos, expertos en derecho, estrategas militares y economistas, respectivamente. Sin

embargo, para conseguir obtener las respuestas que necesitamos para cuando las necesitamos, todo el mundo debe participar en el debate porque, como veremos, las dificultades trascienden todas las fronteras tradicionales, tanto entre disciplinas como entre países.

GAZAPOS FRENTE A IA ROBUSTA

La tecnología de la información ya ha tenido un significativo impacto positivo sobre prácticamente todos los sectores de la actividad humana: desde la ciencia hasta las finanzas, la producción industrial, el transporte, la sanidad, la energía o las comunicaciones. Y este impacto palidece en comparación con el progreso que la IA es capaz de traer. Pero, cuanto más dependamos de la tecnología, más importante será que esta sea robusta y fiable, que haga lo que queremos que haga.

A lo largo de la historia de la humanidad, hemos usado una estrategia muy contrastada para que nuestra tecnología siguiese siendo beneficiosa: aprender de nuestros errores. Inventamos el fuego, cometimos un error tras otro, y más tarde inventamos el extintor, la salida de incendios, la alarma de incendios y el cuerpo de bomberos. Inventamos el automóvil, que sufrió innumerables accidentes, y a continuación inventamos los cinturones de seguridad, los airbags y los coches autónomos. Hasta ahora, nuestras tecnologías han provocado tan pocos accidentes, y de consecuencias tan limitadas, que los beneficios compensaban los daños que causaban. Pero a medida que seguimos desarrollando tecnologías cada vez más potentes, inevitablemente llegará un momento en el que incluso un solo accidente podría ser tan devastador para anular todos los beneficios que conlleva. Hay quien argumenta que una guerra nuclear global accidental podría ser un ejemplo de esta situación. Otros afirman que lo sería una pandemia producto de la bioingeniería, y en el capítulo siguiente abordaremos la controversia sobre si la IA del futuro podría provocar la extinción de la humanidad. Pero no es necesario considerar ejemplos tan extremos para llegar a una conclusión crucial: a medida que la tecnología va siendo cada vez más potente, debemos depender menos de la estrategia de prueba y error para la ingeniería de seguridad. Dicho de otro modo, debemos volvernos más proactivos que

reactivos, e invertir en una investigación en seguridad que ha de ir dirigida a evitar los accidentes. Este es el motivo por el que las sociedades invierten más en la seguridad de los reactores nucleares que en la de las ratoneras.

Esta es también la razón por la que, como vimos en el primer capítulo, los participantes mostraron gran interés por la investigación en IA segura durante la conferencia en Puerto Rico. Desde siempre, los ordenadores y los sistemas de IA se han quedado colgados de vez en cuando, pero esta vez es diferente: la IA va penetrando gradualmente en el mundo real, y, si provoca que dejen de funcionar la red eléctrica, la bolsa o un sistema de armas nucleares, será mucho más que un mero incordio. En el resto de esta sección, quiero presentarle las cuatro áreas principales de investigación técnica que dominan la discusión actual en torno a la IA segura y en las que se trabaja en lugares de todo el mundo: *verificación*, *validación*, *seguridad* y *control*.(9) Para evitar que las cosas se pongan demasiado técnicas y arduas, hagámoslo repasando los éxitos y fracasos pasados de la tecnología de la información en diversas áreas, así como las valiosas lecciones que pueden extraerse de ellos y los retos que plantean a los investigadores.

Aunque la mayoría de estas historias son antiguas, y en ellas aparecen sistemas informáticos poco complejos que casi nadie englobaría dentro de la IA y que provocaron pocas víctimas, si es que las hubo, veremos que nos proporcionan valiosas lecciones para diseñar en el futuro sistemas de IA seguros y potentes, cuyos fallos podrían ser verdaderamente catastróficos.

IA para la exploración espacial

Empecemos por algo que me llega al alma: la exploración espacial. La tecnología de computación nos ha permitido llevar gente a la Luna y enviar astronaves no tripuladas a explorar todos los planetas del sistema solar, e incluso hacer que aterricen en Titán, una de las lunas de Saturno, y en un cometa. Como veremos en el capítulo 6, en el futuro la IA podría ayudarnos a explorar otros sistemas solares y galaxias… si no tiene fallos. El 4 de junio de 1996, los científicos que pretendían investigar la magnetosfera terrestre estallaron en aplausos cuando un cohete Ariane 5 de la Agencia Espacial Europea despegó estruendosamente hacia el espacio con los instrumentos que habían construido. Treinta y siete segundos más tarde, sus sonrisas se

esfumaron cuando el cohete estalló como un espectáculo de fuegos artificiales de cientos de millones de dólares.[16] Se determinó que la causa había sido un fallo de software al manipular un número que era demasiado grande para los 16 bits que se le habían reservado.[17] Dos años después, el Mars Climate Orbiter de la NASA penetró accidentalmente en la atmósfera del planeta rojo y se desintegró porque dos partes distintas del software utilizaban diferentes unidades de fuerza, lo que provocó un error del 445 % en el control de los motores de propulsión del cohete.[18] El segundo gazapo supercostoso de la NASA fue el siguiente: su misión Mariner 1 a Venus explotó tras el lanzamiento desde Cabo Cañaveral el 22 de julio de 1962, después de que el software de control de vuelo fallase debido a un signo de puntuación incorrecto.[19] Como si quisiesen demostrar que los occidentales no eran los únicos que dominaban el arte de lanzar gazapos al espacio, la misión soviética Fobos 1 fracasó el 2 de septiembre de 1988. Se trataba de la astronave interplanetaria más pesada jamás lanzada, con el espectacular objetivo de desplegar una sonda en Fobos, una luna de Marte, pero la misión se desbarató cuando la falta de un guion provocó que se enviase a la nave el comando de «final de misión» cuando esta iba de camino a Marte, lo que hizo que se apagasen todos sus sistemas.[20]

Lo que aprendemos de estos ejemplos es la importancia de lo que los informáticos llaman *verificación*: el proceso de asegurarse de que el software satisface del todo los requisitos que se esperan de él. Cuantas más vidas y recursos hay en juego, mayor es la necesidad de confiar en que el software funcionará como está previsto. Por suerte, la IA puede ayudar a automatizar y mejorar el proceso de verificación. Por ejemplo, no hace mucho se ha podido comprobar matemáticamente que un *kernel* de sistema operativo completo y de propósito general llamado *seL4* cumple con una especificación formal que permite tener una garantía fuerte contra cuelgues y operaciones poco seguras: aunque aún no tiene toda la funcionalidad del Windows de Microsoft o del Mac OS, podemos estar seguros de que no nos sorprenderá con lo que se conoce afectuosamente como «la pantalla azul de la muerte» o «la rueda giratoria de la perdición». La Agencia de Proyectos de Investigación Avanzados de Defensa (DARPA) estadounidense ha respaldado el desarrollo de un conjunto de herramientas de alta garantía de software de código abierto denominadas HACMS (Sistemas Cibermilitares de Alta Garantía, por sus siglas en inglés) que se han demostrado son seguras. Un escollo importante es

hacer que tales herramientas sean lo bastante potentes y fáciles de utilizar para que se generalice su uso. Otra complicación es que la propia tarea de la verificación será cada vez más difícil a medida que el software se traslade a los robots y a nuevos entornos, y que el software tradicional sea reemplazado por sistemas de IA que aprenden continuamente, lo que los lleva a modificar su comportamiento, como vimos en el capítulo 2.

IA para las finanzas

El financiero es otro sector que está siendo transformado por la tecnología de la información, que permite la eficaz reasignación de recursos de un lugar a otro del mundo a la velocidad de la luz, y hace posible la financiación asequible de cualquier cosa desde hipotecas hasta empresas emergentes. Es probable que los progresos en IA ofrezcan en el futuro grandes oportunidades de negocio en torno a las transacciones financieras: hoy en día, la mayoría de las decisiones de compraventa en el mercado bursátil las toman automáticamente ordenadores, y a mis estudiantes que se gradúan en el MIT los tientan muy a menudo con astronómicos salarios de entrada para mejorar las transacciones algorítmicas.

La verificación también es importante para el software financiero, algo que la compañía estadounidense Knight Capital sufrió en carne propia el 1 de agosto de 2012, al perder 440 millones de dólares en cuarenta y cinco minutos al poner en producción un software de gestión de operaciones bursátiles sin verificar.[21] La «Quiebra Instantánea» de un billón de dólares del 6 de mayo de 2010 fue llamativa por otro motivo. Aunque provocó enormes trastornos durante media hora antes de que los mercados se estabilizasen, tiempo en el cual el precio de las acciones de algunas empresas prominentes, como Procter & Gamble, osciló entre un centavo y cien mil dólares,[22] el problema no vino provocado por ningún gazapo o disfunción que la verificación hubiese podido evitar, sino por una violación de las expectativas: los programas de gestión automática de operaciones bursátiles de muchas compañías se encontraron operando en una situación inesperada, en la que sus suposiciones no eran válidas (por ejemplo, la de que si un ordenador de la Bolsa informaba de que una acción tenía un precio de un centavo, ese era el valor real de dicha acción).

La Quiebra Instantánea ejemplifica la importancia de lo que los informáticos llaman *validación*: mientras que la verificación pregunta «¿Construimos el sistema correctamente?», la validación pregunta «¿Construimos el sistema correcto?».(10) Por ejemplo, ¿se basa el sistema en suposiciones que podrían no ser siempre válidas? De ser así, ¿cómo podría mejorarse para que gestionase mejor la incertidumbre?

IA para la producción industrial

Ni que decir tiene que la IA ofrece enormes posibilidades de mejorar la producción industrial al controlar robots que aumentan tanto su eficiencia como su precisión. La continua mejora de las impresoras 3D permite ahora crear prototipos de cualquier cosa, desde edificios de oficinas hasta dispositivos micromecánicos más pequeños que un grano de sal.[23] Mientras que enormes robots industriales son capaces de fabricar coches y aviones, fresadoras, tornos y cizallas y otras herramientas similares controladas por ordenador no solo se usan en las fábricas, sino también por el «movimiento fabricante de base», en el que entusiastas locales materializan sus ideas en más de mil *fab labs* distribuidos por todo el mundo.[24] Pero cuantos más robots tenemos a nuestro alrededor más importante es que verifiquemos y validemos su software. La primera persona de la que hay constancia de que murió por culpa de un robot fue Robert Williams, un trabajador de una fábrica de Ford en Flat Rock (Michigan). En 1979, un robot que debía sacar piezas de una zona de almacenamiento sufrió una avería, y Williams trepó hasta esa zona para coger las piezas. El robot se puso en marcha silenciosamente, aplastó la cabeza de Williams, y siguió funcionando durante treinta minutos hasta que sus compañeros descubrieron lo que había sucedido.[25] La siguiente víctima de un robot fue Kenji Urada, un ingeniero de mantenimiento en una fábrica de Kawasaki en Akashi (Japón). En 1981, mientras reparaba un robot averiado, pulsó accidentalmente el botón de encendido y el brazo hidráulico del robot lo aplastó hasta matarlo.[26] En 2015, un proveedor de una de las fábricas de Volkswagen en Baunatal (Alemania) trabajaba en la puesta a punto de un robot para que este agarrase y manipulase piezas de automóvil cuando algo falló y el robot lo agarró a él y lo aplastó contra una placa de metal hasta matarlo.[27]



FIGURA 3.3. Mientras que los robots industriales tradicionales son caros y difíciles de programar, existe una tendencia hacia robots más baratos y dotados de IA, capaces de aprender de trabajadores sin experiencia en programación.

Aunque estos accidentes son trágicos, es importante señalar que representan una minúscula proporción de todos los accidentes industriales. Más aún, los accidentes industriales han *disminuido* en lugar de aumentar a medida que la tecnología ha ido mejorando, y, en Estados Unidos, las muertes debidas a robots han pasado de unas 14.000 en 1970 a 4.821 en 2014.[28] Los tres accidentes mencionados demuestran que añadir inteligencia a máquinas tontas debería contribuir a mejorar aún más la seguridad industrial, al hacer que los robots aprendiesen a ser más cuidadosos cuando tuviesen personas cerca. Los tres accidentes se podrían haber evitado con una mejor validación: los robots causaron daños no por gazapos o por malicia, sino porque hicieron suposiciones equivocadas: que la persona no estaba allí o que era una pieza de automóvil.

IA para el transporte

Aunque la IA puede salvar vidas en la producción industrial, podría salvar

aún más en el sector del transporte. Solo en accidentes de circulación, murieron más de 1,2 millones de personas en 2015 en todo el mundo, y los accidentes de avión, tren y barco mataron a varios miles más. En Estados Unidos, con sus estrictos estándares de seguridad, los accidentes de vehículos de motor acabaron con la vida de unas 35.000 personas el año pasado (siete veces más que todos los accidentes industriales juntos).[29] Cuando organizamos una mesa redonda sobre este asunto en Austin (Texas) en 2016, dentro del encuentro anual de la Asociación para el Avance de la Inteligencia Artificial, el informático israelí Moshe Vardi habló sobre ello de forma conmovedora y argumentó que la IA no solo podría reducir las víctimas en las carreteras, sino que debía hacerlo: «¡Es un imperativo moral!», exclamó. Como la mayoría de los accidentes de tráfico son debidos a errores humanos, está muy extendida la idea de que los coches autónomos dotados de IA podrían eliminar al menos el 90 % de las muertes en carretera, y este optimismo está impulsando grandes avances en el desarrollo de los coches autónomos. Elon Musk imagina un futuro en el que los coches autónomos no solo serán más seguros, sino que también permitirán que sus dueños ganen dinero con ellos compitiendo con Uber y Lyft cuando no los necesiten.

Hasta ahora, el historial de los coches autónomos en cuanto a seguridad es de hecho mejor que el de los conductores humanos, y los accidentes que se han producido ponen de manifiesto la importancia y la dificultad de la validación. El primer accidente menor causado por un coche autónomo de Google tuvo lugar el 14 de febrero de 2016, porque este hizo una suposición incorrecta sobre un autobús: que su conductor aminoraría la marcha cuando el coche se incorporó delante de él. El primer accidente mortal provocado por un Tesla autónomo, que se estampó contra el remolque de un camión cruzado en una autopista el 7 de mayo de 2016, se debió a dos suposiciones erróneas: [30] que el lateral blanco luminoso del remolque era simplemente parte del cielo soleado, y que el conductor (que al parecer estaba viendo una película de Harry Potter) tenía su atención puesta en la autopista e intervendría si había algún problema.(11)

Pero a veces unas buenas verificación y validación no son suficientes para evitar accidentes, porque también se necesita un buen *control*: la capacidad de que un operador humano supervise el sistema y modifique su comportamiento si fuese necesario. Para que estos sistemas *con intervención humana* funcionen correctamente, es fundamental que la comunicación entre

humano y máquina sea efectiva. Por ejemplo, una luz roja en el salpicadero nos avisará si nos dejamos abierto el maletero del coche. Por desgracia, cuando el *Herald of Free Enterprise*, un ferri británico que transportaba personas y coches, zarpó del puerto de Zeebrugge el 6 de marzo de 1987 con las puertas de su zona de carga abiertas, no disponía de ninguna luz de alerta u otro medio visible de advertencia para el capitán, y el ferri volcó poco después de salir del puerto, causando la muerte de 193 personas.[31]

Otro trágico fallo de control que podría haberse evitado mediante una mejor comunicación entre máquina y humano tuvo lugar durante la noche del 1 de junio de 2009, cuando el vuelo 447 de Air France se estrelló en el océano Atlántico, provocando la muerte de sus 228 ocupantes. Según el informe oficial sobre el accidente, «la tripulación no fue consciente en ningún momento de que el avión estaba perdiendo sustentación, y por lo tanto nunca aplicó una maniobra de recuperación», lo que habría implicado bajar el morro del avión, hasta que fue demasiado tarde. Los expertos en seguridad aérea conjeturaron que el accidente podría haberse evitado si la cabina hubiese dispuesto de un indicador de «ángulo de ataque» que hubiese advertido a los pilotos que el morro estaba demasiado inclinado hacia arriba.[32]

Cuando el vuelo 148 de Air Inter se estrelló contras los montes Vosgos cerca de Estrasburgo, en Francia, el 20 de febrero de 1992, provocando la muerte de 87 personas, la causa no fue la falta de comunicación entre humano y máquina, sino una interfaz de usuario confusa. Los pilotos introdujeron «33» en un teclado porque querían descender con una inclinación de 3,3 grados, pero el piloto automático interpretó esa cifra como 3.300 pies por minuto, porque estaba en un modo distinto, y la pantalla era demasiado pequeña para mostrar el modo y permitir que los pilotos se dieran cuenta de su error.

IA para la energía

La tecnología de la información ha obrado maravillas en relación con la generación y distribución de electricidad, con elaborados algoritmos que equilibran la producción y el consumo en las redes eléctricas mundiales, y complejos sistemas de control que se encargan de que las centrales eléctricas funcionen de manera segura y eficiente. Es probable que el futuro progreso

de la IA haga que la «red inteligente» lo sea aún más, para, idealmente, adaptarse a las variaciones en la oferta y la demanda, incluso hasta llegar al nivel de un panel solar individual en un tejado o una batería para el hogar. Pero el jueves 14 de agosto de 2003 unos 55 millones de personas en Estados Unidos y Canadá se quedaron sin electricidad, y la mayoría de ellos no volvieron a tenerla hasta varios días después. También en este caso se determinó que la causa principal fue un fallo en la comunicación máquina-humano: un fallo de software impidió que el sistema de alarma en una sala de control en Ohio alertase a los técnicos de la necesidad de redistribuir la potencia antes de que un problema menor (que unas líneas de transmisión sobrecargadas estaban chocando con unas ramas sin podar) se escapara de control.[33]

La fusión parcial del núcleo del reactor de Three Mile Island, en Pensilvania, el 28 de marzo de 1979, provocó unos costes de saneamiento de unos mil millones de dólares y el surgimiento de un gran movimiento de rechazo contra la energía nuclear. El informe final del accidente identificó varios factores que contribuyeron al mismo, incluida la confusión provocada por una deficiente interfaz de usuario.[34] En particular, la luz de advertencia que los técnicos pensaban que indicaba si una válvula crítica para la seguridad estaba abierta o cerrada señalaba únicamente si se había enviado la señal de cierre de la válvula, por lo que no supieron que esta se había quedado abierta.

Estos accidentes en los sectores de la energía y del transporte nos enseñan que, al poner a la IA a cargo de un número cada vez mayor de sistemas físicos, debemos dedicar importantes recursos de investigación no solo a hacer que las máquinas funcionen bien por su cuenta, sino a que colaboren de manera efectiva con sus controladores humanos. A medida que la IA se vuelve más inteligente, esto implicará no solo la construcción de buenas interfaces de usuario para compartir la información, sino también determinar cuál es la manera óptima de asignar tareas en los equipos formados por humanos y ordenadores (por ejemplo, identificando situaciones en las que el control debería transferirse), y para aplicar eficientemente el discernimiento humano a las decisiones de mayor impacto, en lugar de distraer a los controladores humanos con un torrente de información sin importancia.

IA para la sanidad

La IA tiene un enorme potencial para mejorar la sanidad. La digitalización de los historiales médicos ya ha permitido a los médicos y a los pacientes tomar decisiones más rápidas y mejores, y recibir ayuda inmediata de expertos de todo el mundo para realizar diagnósticos a partir de imágenes digitales. De hecho, es muy posible que los mejores expertos en la realización de tales diagnósticos sean sistemas de IA, dado el rápido progreso en visión artificial y aprendizaje profundo. Por ejemplo, un estudio holandés de 2015 demostró que el diagnóstico computarizado del cáncer de próstata utilizando imágenes obtenidas por resonancia magnética (MRI, por sus siglas en inglés) era tan bueno como el de los radiólogos humanos,[35] y un estudio realizado en Stanford en 2016 reveló que la IA podía diagnosticar el cáncer de pulmón utilizando imágenes de microscopio mejor incluso que los patólogos humanos.[36] Si el aprendizaje automático puede ayudar a descubrir relaciones entre los genes, las enfermedades y las respuestas a los tratamientos, podría revolucionar la medicina personalizada, hacer que el ganado fuese más sano y posibilitar cultivos más resistentes. Además, los robots podrían llegar a ser cirujanos más precisos y fiables que los humanos, incluso sin usar IA avanzada. En los últimos años, se ha llevado a cabo una amplia variedad de operaciones quirúrgicas robotizadas, que a menudo se han efectuado con mayor precisión y miniaturización y mediante incisiones más pequeñas, conducente todo ello a una menor pérdida de sangre, menos dolor y un periodo de recuperación más corto.

Por desgracia, también en la industria sanitaria se han recibido dolorosas lecciones sobre la importancia de que el software sea potente. Por ejemplo, la máquina de radioterapia Therac-25, de fabricación canadiense, fue diseñada para tratar pacientes de cáncer en dos modos distintos: bien con un haz de electrones de baja potencia o con uno de alta potencia de rayos X de varios megavoltios que se concentraba sobre el objetivo mediante un apantallamiento especial. Desafortunadamente, un software sin verificar y que contenía gazapos provocó que los técnicos administrasen en algunas ocasiones el haz de megavoltios cuando pensaban estar aplicando el haz de baja potencia, y lo hiciesen sin el apantallamiento, lo que acabó cobrándose las vidas de varios pacientes.[37] Muchos más pacientes murieron de sobredosis de radiación en el Instituto Oncológico Nacional de Panamá,

donde aparatos de radioterapia que empleaban cobalto 60 radiactivo se programaron de manera que, en 2000 y 2001, funcionaron durante tiempos de exposición demasiado largos debido a una confusa interfaz de usuario que no había sido debidamente validada.[38] Según un informe reciente,[39] los accidentes en cirugía robotizada están relacionados con 144 muertes y 1.391 lesiones en Estados Unidos entre 2000 y 2013, y entre los problemas comunes están no solo cuestiones relativas al hardware, como las chispas eléctricas o el hecho de que pedazos quemados o rotos de instrumentos cayesen dentro de los pacientes, sino también problemas de software, como movimientos incontrolados o el hecho de que algunos aparatos se apagasen de forma involuntaria.

La buena noticia es que el resto de los casi dos millones de operaciones de cirugía robotizadas de las que el informe daba cuenta transcurrieron sin incidentes, y parece que los robots están haciendo que aumente la seguridad de las operaciones, y no que disminuya. Según un estudio del Gobierno estadounidense, la deficiente atención sanitaria contribuye a más de cien mil muertes anuales solo en Estados Unidos,[40] por lo que el imperativo moral para desarrollar una mejor IA para la medicina es probablemente más determinante que para el caso de los coches autónomos.

IA para la comunicación

En el sector de las comunicaciones es posiblemente donde los ordenadores han tenido hasta la fecha el mayor impacto. Tras la introducción de las centralitas telefónicas computarizadas en los años cincuenta, de internet en los sesenta, y de la World Wide Web en 1989, hoy en día miles de millones de personas se conectan para comunicarse, comprar, leer las noticias, ver películas o jugar a videojuegos, y están acostumbradas a tener toda la información del mundo a un solo clic de distancia, y a menudo gratis. La incipiente *internet de las cosas* promete mejoras en eficiencia, precisión, comodidad y beneficios económicos al conectar todo tipo de objetos, desde lámparas, termostatos y congeladores hasta transpondedores con biochip en animales de ganadería.

Estos espectaculares éxitos a la hora de conectar el mundo han planteado a los informáticos un cuarto reto: deben mejorar no solo la verificación, la

validación y el control, sino también la *seguridad* contra el software malicioso («malware») y los ataques. Mientras que los problemas mencionados anteriormente eran consecuencia de errores accidentales, la seguridad tiene que ver con la mala conducta deliberada. El primer malware que llamó la atención de los medios de comunicación fue el llamado gusano Morris, difundido el 2 de noviembre de 1988, que aprovechaba gazapos existentes en el sistema operativo UNIX. Parece que se trató de un torpe intento de contar cuántos ordenadores estaban conectados a internet y, aunque infectó e hizo que se colgasen alrededor del 10 % de los 60.000 ordenadores que componían internet por aquel entonces, eso no impidió que su creador, Robert Morris, acabase obteniendo un puesto fijo de profesor de informática en el MIT.

Otros malwares aprovechan puntos débiles presentes no en el software sino en las personas. El 5 de mayo de 2000, como si estuviesen celebrando mi cumpleaños, la gente recibió correos electrónicos con el asunto «ILOVEYOU» remitidos por conocidos y colegas, y los usuarios de Microsoft Windows que clicaron en el archivo adjunto «LOVE-LETTER-FOR-YOU.txt.vbs» ejecutaron sin saberlo un *script* que provocó daños en su ordenador y reenvió el mensaje a todos los contactos de sus libretas de direcciones. Creado por dos jóvenes programadores filipinos, este gusano afectó en torno al 10 % de internet, como había sucedido con el gusano Morris, pero, como para entonces internet ya se había extendido de forma considerable, se convirtió en una de las infecciones más grandes de toda la historia, con más de cincuenta millones de ordenadores afectados y unas pérdidas superiores a los cinco mil millones de dólares. Como probablemente haya usted sufrido en carne propia, internet sigue infestada de innumerables clases de malware infeccioso, que los expertos en seguridad clasifican en gusanos, troyanos, virus y otras categorías de nombres pavorosos, y el daño que causan va desde mostrar inocuos mensajes jocosos hasta borrar nuestros ficheros, robar nuestra información personal, espiarnos o secuestrar nuestro ordenador y usarlo para enviar spam.

Mientras que el malware tiene como destino cualquier ordenador que encuentre, los hackers atacan objetivos específicos de su interés (entre los casos recientes que han tenido mucha repercusión están Target, TJ Maxx, Sony Pictures, Ashley Madison, la empresa petrolera saudí Aramco y el Comité Nacional del Partido Demócrata estadounidense). Más aún, los

botines que obtienen parecen cada vez más espectaculares. En 2008, unos hackers robaron 130 millones de números de tarjetas de crédito y otra información relacionada con cuentas de Heartland Payment Systems, mientras que en 2013 obtuvieron los datos de acceso de más de tres mil millones (¡!) de cuentas de correo electrónico de Yahoo.[41] En 2014, los hackers que atacaron la Oficina de Gestión de Personal del Gobierno estadounidense accedieron a información de más de 21 millones de personas, entre las que parece que había empleados con habilitación para acceder a información considerada secreta, así como las huellas dactilares de agentes encubiertos.

En consecuencia, me exaspero cada vez que leo que algún nuevo sistema es al parecer cien por cien seguro e invulnerable al hackeo. Pero es claramente lo que necesitamos que sean los futuros sistemas de IA, «invulnerables al hackeo», antes de ponerlos al mando de sistemas críticos de infraestructuras o de armamento, por poner dos ejemplos, por lo que la importancia creciente de la IA en la sociedad conlleva que aumente en paralelo la importancia de la seguridad informática. Es verdad que algunos ataques se aprovechan de la credulidad humana o de complejas vulnerabilidades en software recién puesto en funcionamiento, pero otros permiten acceder sin autorización a ordenadores remotos, al sacar provecho de sencillos fallos de software que habían pasado desapercibidos durante periodos de tiempo embarazosamente largos. El gazapo conocido como «Heartbleed» se mantuvo de 2012 a 2014 en una de las bibliotecas de software más populares para la comunicación segura entre ordenadores, y el denominado «Bashdoor» estuvo integrado en el propio sistema operativo Unix desde 1989 hasta 2014. Esto significa que las herramientas de IA que mejoren la verificación y la validación mejorarán también la seguridad.

Por desgracia, unos mejores sistemas de IA también pueden usarse para encontrar nuevas vulnerabilidades y llevar a cabo ataques más elaborados. Imaginemos, por ejemplo, que un día recibimos un correo electrónico fraudulento sorprendentemente personalizado que intenta convencernos de que divulguemos información personal. Nos llega desde la dirección de una amiga, enviado por una IA que la ha hackeado y la está suplantando, imitando su estilo al escribir basándose en un análisis de los otros mensajes que nuestra amiga ha enviado, e incluyendo un montón de información personal sobre nosotros obtenida de otras fuentes. ¿Nos dejaríamos engañar?

¿Y si el correo fraudulento procediese en apariencia de la empresa con la que tenemos contratada una tarjeta de crédito y al mensaje le sigue una llamada telefónica de una cordial voz humana que no somos capaces de detectar que está generada por la IA? En la carrera armamentística que tiene lugar hoy en día entre ataque y defensa en el ámbito de la seguridad informática, hasta ahora hay pocos indicios de que la defensa vaya ganando.

LEYES

Los humanos somos animales sociales que sometimos a todas las demás especies y conquistamos la Tierra gracias a nuestra capacidad para cooperar. Hemos desarrollado leyes para incentivar y facilitar la cooperación, por lo que, si la IA es capaz de mejorar nuestros sistemas legales y de gobernanza, puede permitirnos cooperar más eficazmente que nunca, sacando lo mejor de nosotros. En este sentido, hay multitud de oportunidades de mejora, tanto en cómo se aplican las leyes como en la manera en que se elaboran. A continuación exploraremos ambas facetas.

¿Qué es lo primero que le viene a la mente cuando piensa en el sistema judicial de su país? Si son prolongados retrasos, elevados costes y alguna que otra injusticia ocasional, no es usted el único. ¿No sería maravilloso si lo primero en lo que pensase fuese «eficiencia» e «imparcialidad»? Puesto que el proceso legal puede entenderse de manera abstracta como una computación que recibe como información de entrada las evidencias y las leyes y devuelve una decisión, algunos expertos sueñan con automatizarlo por completo mediante *robojueces*: sistemas de IA que aplican incansablemente los mismos elevados estándares legales a cualquier sentencia sin sucumbir a errores humanos como sesgos, fatiga o carencia del conocimiento más actualizado.

Robojueces

Byron De La Beckwith Jr. fue condenado en 1994 por el asesinato en 1963 de Medgar Evers, líder del movimiento por los derechos civiles, pero en Mississippi dos jurados distintos compuestos solo por blancos habían evitado

condenarlo el año posterior al asesinato, aunque las evidencias físicas eran básicamente las mismas.[42] Por desgracia, la historia de la justicia está repleta de sentencias sesgadas debido al color de piel, al género, a la orientación sexual, a la religión, a la nacionalidad y a otros factores. Los robojueces podrían en principio garantizar que, por primera vez en la historia, todas las personas serían de verdad iguales ante la ley: podrían programarse para que fuesen todos idénticos y tratasen a todo el mundo por igual, aplicando la ley con transparencia y de manera imparcial.

Los robojueces podrían eliminar también los sesgos humanos más accidentales que intencionados. Por ejemplo, un controvertido estudio de 2012 sobre los jueces israelíes afirmó que dictaban veredictos bastante más severos cuando tenían hambre: mientras que los jueces denegaban alrededor del 35 % de las peticiones de libertad condicional inmediatamente después del desayuno, esa cifra ascendía hasta el 85 % justo antes del almuerzo.[43] Otro limitación de los jueces humanos es que puede faltarles tiempo suficiente para estudiar todos los detalles de un caso. Por el contrario, los robojueces pueden copiarse con facilidad, ya que son poco más que software, lo que permitiría que todos los casos pendientes se juzgasen en paralelo en lugar de hacerse en serie, y cada caso tendría su propio robojuez durante el tiempo que fuese necesario. Por último, mientras que es imposible que los jueces humanos dominen todo el conocimiento técnico necesario para todos los casos posibles, desde espinosas disputas de patentes hasta asesinatos cuya dilucidación depende de la ciencia forense más novedosa, los futuros robojueces podrían disponer de memoria y capacidad de aprendizaje ilimitadas.

Así pues, algún día estos robojueces podrían ser más eficientes y más justos, gracias al hecho de ser ecuánimes, competentes y transparentes. Su eficiencia los hace aún más justos: al acelerar el proceso legal y dificultar que abogados astutos distorsionasen el resultado, podrían hacer que fuera sustancialmente más barato obtener justicia a través de los tribunales. Esto podría incrementar en gran medida las posibilidades de que individuos o empresas emergentes con recursos limitados pudiesen ganar a un billonario o a una compañía multinacional con un ejército de abogados.

Por otra parte, ¿qué sucedería si los robojueces tuviesen gazapos o fuesen hackeados? Ambas situaciones se han dado ya con las máquinas de votación, y cuando lo que esté en juego sean años de cárcel o millones de dólares en el

banco, los incentivos para que se produzcan ciberataques serán aún mayores. Incluso si la IA es lo bastante potente para que confiemos en que un robojuez utiliza el algoritmo con fidelidad a la ley, ¿sentiría todo el mundo que entiende su razonamiento lógico lo suficiente para respetar su decisión? Esta dificultad se complica aún más por el éxito reciente de las redes neuronales, que a menudo se muestran más eficaces que los algoritmos tradicionales de IA, que son más fáciles de entender, pero a costa de resultar inexplicables. Si los acusados quieren saber por qué han sido condenados, ¿no deberían tener derecho a una respuesta mejor que «entrenamos al sistema usando una enorme cantidad de datos, y esto es lo que decidió»? Además, estudios recientes han demostrado que, si se entrena un sistema de redes neuronales profundas con un gran número de datos de presos, este es capaz de predecir mejor que los jueces humanos quién tiene más probabilidad de volver a delinquir (y por tanto no debería beneficiarse de la libertad condicional). Pero ¿y si el sistema descubre que la reincidencia está ligada estadísticamente al sexo o a la raza del preso? ¿Se consideraría que este juez es sexista o racista y que debería ser reprogramado? De hecho, un estudio de 2016 argumentó que el software para la predicción de la reincidencia que se usaba en todo Estados Unidos estaba sesgado en contra de los afroamericanos y había contribuido a que se dictaran sentencias injustas.[44] Estas son cuestiones importantes sobre las que todos debemos reflexionar y debatir para asegurarnos de que la IA sigue siendo efectivamente útil. La decisión sobre los robojueces no es de todo o nada, sino que tiene que ver con el grado en que queremos desplegar la IA en nuestro sistema judicial, y la velocidad a la que hacerlo. ¿Queremos que los jueces dispongan de sistemas de ayuda a la toma de decisiones basados en IA, como los médicos del mañana? ¿Queremos ir más allá y que los robojueces dicten sentencias que puedan apelarse ante jueces humanos, o queremos llegar hasta el final y cederles también la decisión última a las máquinas, incluso para casos de pena de muerte?

Controversias legales

Hasta ahora, solo nos hemos centrado en la *aplicación* de la ley; fijémonos ahora en su *contenido*. Existe un amplio consenso en torno a que nuestras leyes deben evolucionar para seguir el ritmo que marca la tecnología. Por

ejemplo, los dos programadores que crearon el ya mencionado gusano ILOVEYOU y provocaron daños por valor de miles de millones de dólares fueron absueltos de todos los cargos y puestos en libertad porque, en aquella época, en Filipinas no había leyes contra la creación de malware. Puesto que la velocidad del progreso tecnológico parece estar acelerándose, hay que actualizar las leyes cada vez más rápido, y tienden a quedar rezagadas. Sin duda, a la sociedad le convendría que en las facultades de derecho y en los gobiernos entrasen personas mucho más duchas en tecnología. Pero ¿debería ir eso seguido de sistemas de ayuda a la toma de decisiones basados en IA para votantes y legisladores, y a continuación directamente de robolegisladores?

Cuál es la mejor manera de modificar nuestras leyes para reflejar el progreso de la IA es un fascinante tema de discusión. En particular, hay una controversia que refleja la tensión existente entre privacidad y libertad de información. Los defensores de la libertad argumentan que, cuanto menor sea nuestra privacidad, con más evidencias contarán los tribunales y más justas serán las sentencias. Por ejemplo, si el Gobierno espiase los dispositivos electrónicos de todos nosotros para tener constancia de dónde estamos, qué escribimos, clicamos, decimos y hacemos, muchos delitos se revolverían fácilmente, y otros se evitarían. Los defensores de la privacidad replican diciendo que no quieren que se instaure un Estado de vigilancia orwelliano, y, en el caso de desearlo, existe el riesgo de que se convierta en una dictadura totalitaria de proporciones épicas. Además, las técnicas de aprendizaje automático han mejorado mucho en cuanto al análisis de datos cerebrales obtenidos mediante escáneres de IRMf para determinar en qué está pensando una persona y, en particular, si dice la verdad o miente.[45] Si la tecnología de escaneo cerebral con ayuda de IA llegase a ser algo habitual en los tribunales, el proceso actualmente laborioso de establecer los hechos de un caso se podría simplificar y acelerar de forma drástica, lo que haría posibles juicios más rápidos y justos. Pero los defensores de la privacidad sin duda temerían que estos sistemas pudiesen cometer errores y, lo que es más importante, que nuestras mentes deberían ser terreno vedado para el espionaje gubernamental. Los gobiernos que no defienden la libertad de pensamiento podrían utilizar esa tecnología para criminalizar a quienes tuvieran ciertas creencias y opiniones. ¿Dónde trazaría usted la línea entre la justicia y la privacidad, y entre proteger la sociedad y hacer lo propio con la libertad

personal? Con independencia de dónde trazase la línea, ¿se desplazaría gradual pero inexorablemente hacia una menor privacidad para compensar el hecho de que la evidencia es más fácil de falsificar? Por ejemplo, una vez que la IA sea capaz de generar vídeos falsos plenamente realistas de usted cometiendo un delito, ¿votaría a favor de un sistema en el que el Gobierno guardase información sobre la ubicación de todas las personas en todo momento y pudiese proporcionarle una coartada irrefutable si fuese necesario?

Otra controversia fascinante es la que gira en torno a si la investigación en IA debería regularse o, de manera más general, qué incentivos debería dar la administración a los investigadores en IA para maximizar la probabilidad de que los resultados de dicha investigación sean útiles. Algunos investigadores se han manifestado en contra de toda forma de regulación del desarrollo de la IA, afirmando que esta retrasaría innecesariamente innovaciones de las que hay una necesidad urgente (por ejemplo, coches autónomos capaces de salvar vidas) y empujaría la investigación puntera en IA hacia la clandestinidad y/o a otros países con gobiernos más permisivos. En la conferencia sobre IA útil organizada en Puerto Rico que se mencionó en el primer capítulo, Elon Musk argumentó que lo que necesitamos ahora mismo de los gobiernos no es supervisión sino perspicacia: en particular, personas técnicamente competentes en puestos de la administración capaces de vigilar el progreso de la IA y dirigirlo si en algún momento fuese necesario. También argumentó que la regulación gubernamental puede en ocasiones fomentar el progreso, en lugar de entorpecerlo: por ejemplo, si los estándares públicos de seguridad para los coches autónomos pueden contribuir a la reducción del número de accidentes en los que estos se vean envueltos, eso haría menos probable el rechazo público, lo cual podría acelerar la adopción de la nueva tecnología. Las empresas con una mentalidad más orientada hacia la seguridad podrían ver con buenos ojos una regulación que obligaría a sus competidores menos escrupulosos a satisfacer sus estándares más estrictos.

Otra interesante controversia legal tiene que ver con la concesión de derechos a las máquinas. Si los coches autónomos redujesen a la mitad las 32.000 víctimas anuales de accidentes de tráfico en Estados Unidos, puede que los fabricantes no recibieran 16.000 notas de agradecimiento, sino 16.000 demandas. Si un coche autónomo provoca un accidente, ¿quién debería ser responsable: sus ocupantes, su dueño o su fabricante? El jurista David

Vladeck ha propuesto una cuarta posibilidad: ¡el propio coche! En particular, Vladeck propone que se permita que los coches autónomos tengan un seguro de automóvil (y se los obligue a hacerlo). De esta manera, los modelos con un historial de seguridad intachable podrían tener derecho a primas muy bajas, probablemente inferiores a las que pagan los conductores humanos, mientras que los modelos con diseños deficientes de fabricantes chapuceros solo podrían contratar pólizas cuyo coste resultaría prohibitivo.

Pero si máquinas como los coches pudiesen contratar pólizas de seguros, ¿deberían también tener la posibilidad de poseer dinero y propiedades? De ser así, legalmente nada impediría que ordenadores inteligentes ganasen dinero en bolsa y lo usasen para costear servicios online. Una vez que un ordenador empieza a pagar a humanos para que trabajen para él, puede conseguir todo aquello de lo que los humanos son capaces. Si la capacidad para la gestión de inversiones de los sistemas de IA llega alguna vez a superar la de los humanos (cosa que ya sucede en algunos ámbitos), esto podría conducir a una situación en la que las máquinas poseyesen y controlasen la mayor parte de nuestra economía. ¿Es eso lo que queremos? Si esto nos parece algo remoto, recordemos que la mayor parte de nuestra economía ya es propiedad de otras entidades no humanas: las corporaciones, que muchas veces son más poderosas que cualquier persona de las que trabajan en ellas y pueden, en cierta medida, adquirir vida propia.

Si nos parece bien que se les conceda a las máquinas el derecho de tener propiedades, ¿qué nos parecería que tuviesen derecho de voto? Si también se les concediese, ¿debería cada programa tener un voto, aunque este pudiese fácilmente hacer billones de copias de sí mismo en la nube si tuviese dinero suficiente, asegurándose así un papel decisivo en las elecciones? En caso contrario, ¿cuál sería la justificación moral para discriminar a las mentes artificiales frente a las humanas? ¿Es relevante el hecho de que esas mentes artificiales sean o no conscientes en el sentido de tener una experiencia subjetiva como nosotros? Estudiaremos en más profundidad estas controvertidas cuestiones relacionadas con el control de nuestro mundo por los ordenadores en el siguiente capítulo, y las relacionadas con la consciencia de las máquinas en el capítulo 8.

ARMAS

Desde tiempo inmemorial, la humanidad ha padecido hambrunas, enfermedades y guerras. Ya hemos mencionado cómo puede la IA contribuir a reducir las dos primeras; y ¿qué hay de las guerras? Hay quien argumenta que el hecho de que las armas nucleares sean tan terroríficas disuade a los países de entrar en guerra. ¿Y si permitiésemos que todos los países construyesen armas basadas en IA cada vez más terribles con la esperanza de que eso acabase definitivamente con todas las guerras? Si este argumento no le convence y cree que es inevitable que haya guerras en el futuro, ¿qué le parecería usar la IA para que estas fuesen menos crueles? Si las guerras consistiesen tan solo en combates de máquinas contra máquinas, no tendría por qué morir ningún ser humano, soldado o civil. Además, cabe la esperanza de que en el futuro los drones dotados de IA y de otros sistemas de armamento robótico (AWS, por sus siglas en inglés; también conocidos por sus detractores como «robots asesinos») podrían ser más justos y racionales que los soldados humanos: equipados con sensores sobrehumanos y sin miedo a morir, mantendrían la calma y se comportarían de manera calculadora y ecuánime incluso en el fragor de la batalla, y sería menos probable que matasen civiles.



FIGURA 3.4. Mientras que los drones militares actuales (como el MQ-1 Predator de la Fuerza Aérea estadounidense) los manejan humanos por control remoto, los drones dotados de IA del futuro podrían

hacer innecesaria toda intervención humana, usando un algoritmo para elegir sus objetivos y a quién matar.

Sistemas con intervención humana

Pero ¿qué pasaría si los sistemas automatizados son defectuosos, o confusos o no tienen el comportamiento esperado? El sistema estadounidense Phalanx para los cruceros de clase Aegis automáticamente detecta, rastrea y neutraliza amenazas como misiles antibuque y aviones. El USS *Vincennes* era un crucero dotado de misiles guiados, apodado Robocruiser por su sistema Aegis. El 3 de julio de 1988, en el fragor de una escaramuza con cañoneras iraníes durante la guerra entre Irán e Irak, su sistema de radar avisó de que se aproximaba un avión. El capitán, William Rodgers III, supuso que estaban siendo atacados por un caza F-14 iraní lanzado en picado y dio al sistema Aegis permiso para disparar. Lo que no supo en ese momento era que había derribado el avión civil de pasajeros que hacía el vuelo 655 de Iran Air, acabando con la vida de sus 290 tripulantes y provocando la indignación internacional. La posterior investigación sacó a la luz una interfaz de usuario confusa que no mostraba automáticamente qué puntos en la pantalla del radar correspondían a aviones civiles (el vuelo 655 seguía su recorrido diario habitual y llevaba encendido el transpondedor que lo identificaba como avión civil) o qué puntos estaban descendiendo (para lanzar un ataque) o ascendiendo (que es lo que el vuelo 655 estaba haciendo tras haber despegado del aeropuerto de Teherán). Por el contrario, cuando al sistema automatizado se le pidió información sobre el avión misterioso, respondió que estaba «descendiendo» porque ese era el estado de otro avión distinto al que, de manera confusa, había reasignado un número que la Marina usaba para hacer seguimiento de los aviones: el que descendía era un avión estadounidense que realizaba una patrulla aérea de combate de superficie muy lejos, en el golfo de Omán.

En este ejemplo, un humano intervenía para tomar la decisión final y, bajo la presión de actuar con rapidez, confió demasiado en lo que le decía el sistema automatizado. Hasta ahora, según los ministerios de Defensa de los distintos países, todos los sistemas de armamento cuentan con intervención humana, a excepción de trampas poco sofisticadas como las minas terrestres.

Pero hoy en día se están desarrollando armas totalmente autónomas que seleccionan y atacan objetivos de manera independiente. Desde un punto de vista militar, resulta tentador eliminar cualquier intervención humana para ganar rapidez: en una refriega entre un dron del todo autónomo capaz de reaccionar al instante y otro que lo hace más despacio porque lo controla remotamente un humano situado en la otra punta del mundo, ¿cuál cree usted que se impondrá?

Sin embargo, ha habido situaciones muy comprometidas en las que hemos tenido suerte de que hubiese una intervención humana. El 27 de octubre de 1962, durante la crisis de los misiles cubanos, once destructores de la Marina estadounidense y el portaaviones USS *Randolph* habían arrinconado al submarino soviético B-59 cerca de Cuba, en aguas internacionales fuera del área de «cuarentena» estadounidense. Lo que no sabían es que la temperatura a bordo había sobrepasado los 45 grados centígrados, porque las baterías del submarino se estaban agotando y el aire acondicionado había dejado de funcionar. Al borde del envenenamiento por dióxido de carbono, muchos miembros de la tripulación habían perdido el conocimiento. Hacía varios días que la tripulación no había tenido contacto con Moscú y no sabía si había estallado la Tercera Guerra Mundial. Fue entonces cuando los estadounidenses empezaron a lanzar pequeñas cargas de profundidad para que el submarino emergiese a la superficie y abandonase la zona, según le habían comunicado a Moscú sin que la tripulación del submarino estuviese al tanto de ello. «Pensamos: se acabó, es el fin —recordaba V. P. Orlov, uno de los tripulantes—. Era como si estuviésemos metidos en un barril metálico y alguien no dejase de darle martillazos.» Lo que los estadounidenses tampoco sabían era que la tripulación del B-59 tenía un torpedo nuclear y la autorización para lanzarlo sin necesidad de pedir permiso a Moscú. De hecho, eso fue lo que decidió hacer el capitán Savitski. Valentín Grigoriévich, el oficial encargado de los torpedos, exclamó: «¡Moriremos, pero los hundiremos a todos, no deshonraremos a nuestra marina!». Afortunadamente, la decisión de lanzar el torpedo debía ser autorizada por tres oficiales a bordo del submarino, y uno de ellos, Vasili Arjípov, se negó a hacerlo. Da que pensar el hecho de que muy poca gente haya oído hablar de Arjípov, pese a que su decisión quizá evitó la Tercera Guerra Mundial y puede que haya sido la contribución individual a la humanidad más importante en toda la historia moderna.[46] También resulta aleccionador imaginar lo que podría haber

sucedido si el B-59 hubiera sido un submarino autónomo controlado por IA sin ninguna intervención humana.

Dos décadas más tarde, el 9 de septiembre de 1983, volvió a elevarse la tensión entre las dos superpotencias: poco tiempo antes, el presidente estadounidense Ronald Reagan había calificado a la Unión Soviética de «imperio del mal» y, apenas una semana antes, la URSS había derribado un avión de pasajeros de Korean Airlines que había penetrado en su espacio aéreo, matando a 269 personas, incluido un congresista estadounidense. Ese día, un sistema automatizado de alerta temprana soviético avisó de que Estados Unidos había lanzado cinco misiles nucleares tierra-tierra hacia la Unión Soviética, y el oficial Stanislav Petrov dispuso solo de unos minutos para decidir si se trataba de una falsa alarma. Se comprobó que el satélite funcionaba correctamente, por lo que, según el protocolo, debería haber informado de que se estaba produciendo un ataque nuclear. Sin embargo, fiándose de su instinto y pensando que era poco probable que Estados Unidos atacase solo con cinco misiles, informó a sus superiores de que era una falsa alarma sin saber si era así. Más tarde se supo que un satélite había tomado los reflejos de la luz solar sobre la parte superior de las nubes por llamaradas de los motores de los cohetes.[47] Me pregunto qué habría pasado si en lugar de Petrov hubiese habido un sistema de IA que siguiese debidamente el protocolo adecuado.

¿La próxima carrera armamentística?

Como sin duda a estas alturas ya habrá imaginado, personalmente tengo serias preocupaciones sobre los sistemas armamentísticos autónomos. Pero ni siquiera he empezado aún a explicarle cuál es mi mayor preocupación al respecto: el desenlace de una carrera armamentística en armas dotadas de IA. En julio de 2015, expresé esta preocupación a través de la siguiente carta abierta que firmé junto a Stuart Russell, con útiles contribuciones de mis colegas del Future of Life Institute:[48]

ARMAS AUTÓNOMAS

Carta abierta de investigadores en IA y robótica

Las armas autónomas seleccionan objetivos y lanzan ataques contra ellos sin intervención humana. Esta categoría de armas incluye, por ejemplo, los quadcópteros armados capaces de rastrear a personas que cumplan determinados criterios predefinidos y eliminarlas, pero no así los misiles de crucero o los drones pilotados remotamente, en los que son los humanos quienes toman todas las decisiones en relación a sus objetivos. La tecnología de la inteligencia artificial (IA) ha alcanzado un grado de desarrollo, en el cual el despliegue de ese tipo de sistemas será factible en la práctica (aunque no lo sea desde el punto de vista legal) en unos cuantos años, no décadas, y es mucho lo que está en juego: las armas autónomas se consideran la tercera revolución en las guerras, tras la pólvora y las armas nucleares.

Se han expuesto muchos argumentos a favor y en contra de las armas autónomas. Por ejemplo, que reemplazar los soldados humanos por máquinas tendrá el efecto positivo de reducir el número de bajas que sufrirá quien las posea, pero también la consecuencia negativa de rebajar el umbral a sobrepasar para decidir entrar en combate. La cuestión clave para la humanidad hoy en día pasa por decidir si iniciamos una carrera armamentística global en IA o evitamos que esto suceda. Si alguna gran potencia militar sigue adelante con el desarrollo de armamento dotado de IA, una carrera armamentística global será prácticamente inevitable, y el resultado final de esa trayectoria tecnológica es evidente: las armas autónomas serán los kaláshnikovs del futuro. A diferencia de las armas nucleares, no requieren materias primas costosas o difíciles de conseguir, por lo que resultarán lo bastante ubicuas y baratas para que todas las potencias militares importantes las produzcan en masa. Será solo cuestión de tiempo que aparezcan en el mercado negro y en manos de terroristas, dictadores que aspiren a reafirmar su control sobre sus poblaciones, señores de la guerra que busquen perpetrar limpiezas étnicas, etcétera. Las armas autónomas son ideales para llevar a cabo asesinatos, desestabilizar países, someter a poblaciones enteras o cometer matanzas selectivas de determinados grupos étnicos. Creemos por tanto que una carrera armamentística de IA militar no sería beneficiosa para la humanidad. Hay muchas formas en las que la IA puede hacer que los campos de batalla resulten más seguros para los humanos, especialmente para los civiles, sin necesidad de crear nuevas herramientas para matar personas.

De la misma manera en que la mayoría de los químicos y los biólogos no tienen ningún interés en producir armas químicas o biológicas, gran parte de los investigadores en IA no tenemos ningún interés en fabricar armas dotadas de IA ni queremos que otros ensucien nuestro campo al hacerlo, con el riesgo añadido de dar pie a una reacción contraria a la IA que restrinja los beneficios que esta podría proporcionar a la sociedad en el futuro. De hecho, tanto los químicos como los biólogos han apoyado mayoritariamente los acuerdos internacionales que han logrado prohibir las armas químicas y biológicas, del mismo modo en que los físicos apoyaron los tratados que prohibían las armas nucleares en el espacio o las armas láser cegadoras.

Para que fuese más difícil desestimar nuestra preocupación por proceder simplemente de pacifistas ecologistas, intenté que firmase nuestra carta el mayor número posible de investigadores en IA y expertos en robótica profesionales. Con anterioridad, la Campaña Internacional por el Control de las Armas Robóticas había recogido cientos de firmas a favor de la prohibición de los robots asesinos, y yo sospechaba que nosotros podríamos

conseguir aún más. Sabía que las organizaciones profesionales serían reacias a compartir las extensas listas con los correos electrónicos de sus miembros para un fin que podía interpretarse como político, así que recopilé varias listas de nombres e instituciones de investigadores a partir de documentos que encontré online y lancé una campaña para encontrar sus direcciones de correo electrónico en MTurk (Mechanical Turk), la plataforma de microfinanciación colectiva de Amazon. Las direcciones de la mayoría de los investigadores figuran en los sitios webs de sus universidades, de manera que, veinticuatro horas y 54 dólares más tarde, tenía en mi poder una lista de correo de cientos de investigadores en IA que habían destacado lo suficiente en sus carreras para ser elegidos miembros de la Asociación para el Avance de la Inteligencia Artificial (AAAI, por sus siglas en inglés). Uno de ellos era Toby Walsh, profesor de IA británico-australiano, que se comprometió amablemente a escribir al resto de integrantes de la lista y ayudó a encabezar nuestra campaña. Los trabajadores de MTurk de distintos lugares del mundo produjeron otras listas de correo para Toby, y al cabo de poco tiempo más de 3.000 investigadores en IA y robótica habían firmado nuestra carta abierta, entre ellos seis antiguos presidentes de la AAAI y líderes de la industria de la IA de Google, Facebook, Microsoft y Tesla. Un ejército de voluntarios trabajó sin descanso para validar la lista de signatarios, eliminando los nombres falsos como Bill Clinton o Sarah Connor. También firmaron otras 17.000 personas más, entre ellas Stephen Hawking, y, cuando Toby convocó una rueda de prensa sobre la carta en la Conferencia Internacional Conjunta sobre Inteligencia Artificial, fue noticia en todo el mundo.

Como los biólogos y los químicos habían alzado la voz en su momento, sus respectivos campos ahora son conocidos principalmente por crear medicamentos y materiales beneficiosos, en lugar de armas biológicas y químicas. Las comunidades de la IA y la robótica se han manifestado de la misma manera: los firmantes de la carta también querían que sus campos fuesen conocidos por crear un futuro mejor, no por inventar nuevas maneras de matar gente. Pero, en el futuro, ¿el uso de la IA será principalmente civil o militar? Aunque hemos dedicado más páginas a este capítulo que al primero, puede que pronto estemos dedicando más dinero al segundo, en particular si se desencadena una carrera armamentística en la IA militar. La inversión dedicada a IA civil superó los mil millones de dólares en 2016, pero esta cifra palidece ante la petición presupuestaria del Pentágono de entre 12.000 y

15.000 millones de dólares para proyectos relacionados con la IA, y es probable que China y Rusia tomen buena nota de las palabras del vicesecretario de Defensa estadounidense Robert Work cuando lo anunció: «Queremos que nuestros competidores se pregunten qué hay entre bambalinas».[49]

¿Debería firmarse un tratado internacional?

Aunque actualmente hay una importante corriente internacional favorable a la negociación de alguna forma de prohibición de los robots asesinos, aún no está claro qué sucederá, y está teniendo lugar un vibrante debate sobre qué debería suceder. Aunque muchos de los principales actores coinciden en que las potencias mundiales deberían elaborar alguna clase de regulación internacional para orientar la investigación y el uso de los AWS, el acuerdo es menor sobre qué armas en concreto deberían prohibirse y cómo debería implantarse la prohibición. Por ejemplo, ¿deberían prohibirse únicamente las armas autónomas letales, o también las que provocan lesiones graves en las personas, por ejemplo al dejarlas ciegas? ¿Prohibiríamos su desarrollo, su producción o su posesión? ¿Debería la prohibición extenderse a todos los sistemas de armas autónomos o, como decía nuestra carta, solo a los ofensivos, y permitir los sistemas defensivos tales como baterías antiaéreas y defensas antimisiles autónomas? En este último caso, debería un AWS considerarse defensivo aunque pudiese introducirse fácilmente en territorio enemigo? ¿Cómo se aplicaría dicho tratado habida cuenta de que la mayoría de los componentes de un arma autónoma tienen también un uso civil dual? Por ejemplo, no hay mucha diferencia entre un dron capaz de repartir paquetes de Amazon y otro que pueda lanzar bombas.

Hay quien argumenta que diseñar un tratado sobre AWS efectivo es una tarea abocada al fracaso y que por ello no deberíamos ni siquiera intentarlo. Por otra parte, cuando anunció las misiones a la Luna, John F. Kennedy hizo hincapié precisamente en que merece la pena esforzarse por conseguir metas difíciles cuando el éxito al alcanzarlas podría ser muy beneficioso para el futuro de la humanidad. Además, muchos expertos explican que las prohibiciones de las armas biológicas y químicas fueron útiles, aunque su aplicación en la práctica resultó ser difícil, y el grado de incumplimiento,

significativo, ya que las prohibiciones provocaron una fuerte estigmatización de dichas armas que limitó su uso.

Conocí a Henry Kissinger en una cena en 2016, y tuve ocasión de preguntarle sobre su papel en la prohibición de armas biológicas. Me explicó cómo, cuando era consejero de Seguridad Nacional, convenció al presidente Nixon de que una prohibición sería positiva para la seguridad nacional de Estados Unidos. Me impresionó lo agudas que eran su mente y su memoria para tener noventa y dos años, y me resultó fascinante poder escuchar cómo lo veía él desde dentro. Puesto que Estados Unidos ya había alcanzado un estatus de superpotencia gracias a sus fuerzas convencionales y nucleares, era más lo que podía perder que lo que ganaría con una carrera armamentística mundial de armas biológicas de resultado incierto. Dicho de otro modo, si uno ocupa ya la primera posición, tiene sentido guiarse por la sentencia según la cual «si no está roto, para qué arreglarlo». Stuart Russell se sumó a nuestra conversación de sobremesa tras la cena, y discutimos cómo se puede aplicar exactamente el mismo argumento a las armas autónomas letales: quienes más pueden ganar con una carrera armamentística no son las superpotencias, sino pequeños estados rebeldes y actores no estatales como los grupos terroristas, que, una vez que las armas se hubiesen desarrollado, podrían tener acceso a ellas a través del mercado negro.

En cuanto se produjese en masa, un pequeño dron asesino dotado de IA no costaría mucho más que un teléfono inteligente. Ya fuese un terrorista decidido a asesinar a un político o un amante despechado deseando vengarse de su exnovia, no tendrían más que cargar la foto y la dirección de su víctima en el dron asesino, y este podría volar hasta su destino, identificar y eliminar a la persona, y autodestruirse para garantizar que nadie supiese quién había sido el responsable del asesinato. Otra posibilidad, para aquellos decididos a llevar a cabo una limpieza étnica, es que el dron pudiera programarse fácilmente para matar a personas de un determinado color de piel u origen étnico. Stuart imagina que cuanto más inteligentes fuesen esas armas, menos materiales, potencia de fuego y dinero serían necesarios para cada muerte. Por ejemplo, él teme que se creen drones del tamaño de un abejorro que maten por muy poco dinero y que usen una potencia explosiva mínima al disparar a sus víctimas en los ojos, que son lo bastante blandos para permitir que incluso un proyectil pequeño llegase hasta el cerebro. O podrían aferrarse a la cabeza con unas garras metálicas, y a continuación perforar el cerebro

usando una diminuta carga explosiva. Si en un solo camión cupiesen un millón de estos drones asesinos, tendríamos una terrorífica arma de destrucción masiva de una clase completamente nueva: capaz de matar de forma selectiva a una categoría de personas preestablecida, sin hacer ningún daño a las demás.

Un contraargumento habitual es que podemos eliminar este temor haciendo que los robots asesinos tengan cierta ética (por ejemplo, que maten solo a soldados enemigos). Pero si lo que nos preocupa es la aplicación de una prohibición, entonces ¿cómo podría resultar más fácil hacer cumplir el requisito de que las armas autónomas enemigas sean cien por cien éticas que impedir directamente que se produzcan? ¿Se puede afirmar que a los soldados con una excelente formación de los países civilizados se les da tan mal cumplir las reglas de la guerra que los robots lo harían sin duda mejor, y decir al mismo tiempo que los países rebeldes, dictadores y grupos terroristas se ajustan tan escrupulosamente a las normas de la guerra que nunca optarían por desplegar robots violando así esas reglas?

Ciberguerra

Otra faceta militar interesante de la IA es que puede permitirnos atacar al enemigo sin necesidad de fabricar arma alguna, solo a través de la guerra informática. Como pequeño preludio de lo que el futuro podría depararnos, el gusano Stuxnet, cuya autoría se atribuye generalmente a los gobiernos estadounidense e israelí, infectó las ultracentrifugadoras vinculadas al programa iraní de enriquecimiento de combustible nuclear y provocó su autodestrucción. Cuanto mayor sea el grado de automatización de una sociedad, y más potente la IA atacante, más destructiva puede ser la guerra informática. Si somos capaces de hackear los vehículos autónomos, aviones autopilotados, reactores nucleares, robots industriales, sistemas de comunicaciones, sistemas financieros y redes eléctricas del enemigo y hacer que dejen de funcionar, podemos de hecho hundir su economía e inutilizar sus defensas. Si además encontramos la manera de hackear algunos de sus sistemas de armamento, mejor aún.

Comenzamos este capítulo repasando lo espectaculares que son las oportunidades a corto plazo de que la IA beneficie a la humanidad, siempre

que consigamos hacer que sea robusta e invulnerable al hackeo. Aunque se puede usar la propia IA para hacer que los sistemas de IA sean más robustos, contribuyendo así a la defensa en la ciberguerra, evidentemente la IA también puede ayudar al bando atacante. Asegurarse de que la defensa se impone al ataque debe ser uno de los objetivos fundamentales a corto plazo en el desarrollo de la IA, pues de lo contrario todas las fantásticas tecnologías que construyamos podrían volverse en nuestra contra.

TRABAJOS Y SALARIOS

Hasta ahora, a lo largo de este capítulo nos hemos centrado principalmente en cómo la IA nos afectará en cuanto consumidores, al hacer posible que nuestros productos y servicios innovadores se adquieran a precios asequibles. Pero ¿de qué manera nos afectará como trabajadores, al transformar el mercado de trabajo? Si podemos encontrar la forma de incrementar nuestra prosperidad mediante la automatización sin que haya gente que quede privada de ingresos o de un propósito en la vida, entonces podremos crear un futuro fantástico con ocio y una inusitada opulencia para todo el que la quiera. Pocas personas han reflexionado sobre esto más largo y tendido que Erik Brynjolfsson, uno de mis colegas del MIT. Aunque siempre va bien arreglado y viste impecablemente, tiene ascendencia islandesa, y a veces no puedo evitar imaginar que acaba de recortarse sus indómitas barba y melena pelirrojas vikingas para pasar desapercibido en nuestra escuela de negocios. Lo que sin duda no se ha recortado son sus excéntricas ideas. Erik se refiere a su visión optimista del mercado laboral como la «Atenas digital». La razón principal por la cual los ciudadanos atenienses en la Antigüedad llevaban vidas ociosas y podían disfrutar de la democracia, el arte y los juegos era que tenían esclavos que realizaban gran parte del trabajo. Pero ¿por qué no sustituir a los esclavos por robots dotados de IA y crear una utopía digital de la que pueda disfrutar todo el mundo? La economía basada en la IA que Erik imagina no solo eliminaría el estrés y los trabajos pesados, y generaría en abundancia todas las cosas que hoy en día deseamos, sino que también proporcionaría infinitos y maravillosos nuevos productos y servicios que los consumidores actuales ni siquiera saben que quieren.

Tecnología y desigualdad

Podemos llegar desde el punto en el que estamos hoy en día hasta la Atenas digital de Erik si el sueldo por hora de todo el mundo sigue creciendo año a año, de modo que quienes quieran disfrutar de más tiempo de ocio puedan trabajar cada vez menos, mientras mejoran su nivel de vida. La figura 3.5 muestra que esto es exactamente lo que sucedió en Estados Unidos desde la Segunda Guerra Mundial hasta mediados de la década de 1970: aunque había desigualdad económica, el tamaño total del pastel creció de tal manera que también lo hizo el pedazo que recibía casi todo el mundo. Pero entonces, como Erik es el primero en reconocer, algo cambió: la figura 3.5 muestra que, aunque la economía siguió creciendo y eso hizo que también creciese la renta media, las ganancias conseguidas a lo largo de las últimas cuatro décadas fueron a parar a los más adinerados, principalmente al 1 % más rico, mientras que el 90 % más pobre vio cómo sus ingresos se estancaban. El consiguiente aumento de la desigualdad resulta aún más patente si en lugar de fijarnos en la renta lo hacemos en la riqueza. El patrimonio neto medio del 90 % más pobre de los hogares estadounidenses era en 2012 de 85.000 dólares —el mismo que veinticinco años antes—, mientras que, durante ese mismo periodo y teniendo en consideración la inflación, el 1 % más rico había más que doblado su patrimonio, hasta llegar a los 14 millones de dólares.[50] Las diferencias son aún más acusadas a escala internacional: en 2013, la riqueza total de la mitad más pobre de la población mundial (más de 3.600 millones de personas) era igual a la de las ocho personas más ricas del mundo[51]; una estadística que pone de manifiesto tanto la pobreza y vulnerabilidad de los más desfavorecidos, como la opulencia de los más ricos. En nuestra conferencia de 2015 en Puerto Rico, Erik les contó a los investigadores en IA allí reunidos que pensaba que los avances en IA y en automatización seguirían haciendo que creciese el pastel económico, pero que no existía ninguna ley económica según la cual todo el mundo, o ni siquiera la mayoría de las personas, fuese a beneficiarse de ello.

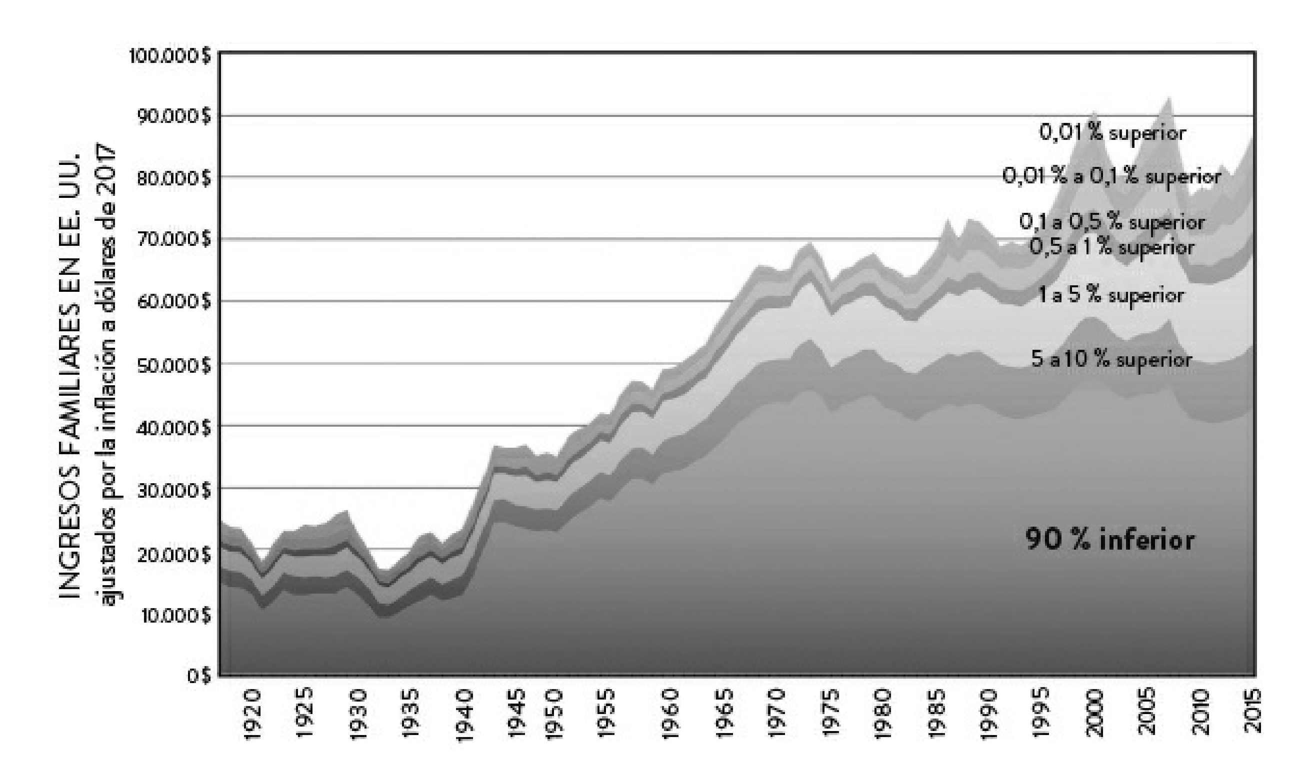


FIGURA 3.5. Cómo la economía ha hecho que aumentase la renta media a lo largo del siglo pasado, y qué proporción de esta renta ha ido a los distintos grupos. Antes de la década de 1970, se puede ver cómo los ricos y los pobres mejoraban su situación al unísono; desde entonces, la mayor parte de las ganancias han ido a parar al 1 % superior, mientras que el 90 % inferior en promedio prácticamente no ha ganado nada.[52] Las cantidades se han ajustado por la inflación a dólares de 2017.

Aunque hay un amplio consenso entre los economistas sobre el aumento de la desigualdad, existe una interesante polémica sobre por qué es así y si esta tendencia continuará. Quienes discuten desde la izquierda del espectro político suelen argumentar que la causa principal es la globalización y/o políticas económicas, como las rebajas de impuestos para los ricos. Pero Erik Brynjolfsson y su colaborador Andrew McAfee, también del MIT, explican que la causa primordial es otra: la tecnología.[53] En particular, argumentan que la tecnología digital exacerba la desigualdad de tres maneras distintas.

En primer lugar, al reemplazar trabajos antiguos por otros que requieren mayor formación, la tecnología premia a quienes tienen más educación: desde mediados de la década de 1970, los salarios de quienes poseían un título universitario aumentaron en torno al 25 %, mientras que aquellos que no habían completado la educación secundaria vieron cómo sus salarios se reducían en promedio en un 30 %.[54]

En segundo lugar, afirman que, desde el año 2000, una proporción cada

vez mayor de la renta empresarial ha ido a parar a los dueños de las compañías, en detrimento de quienes trabajan en ellas, y que, mientras continúe la automatización, deberíamos esperar que los propietarios de las máquinas se queden con un pedazo cada vez más grande del pastel. Esta ventaja del capital sobre la mano de obra puede ser particularmente importante para la creciente economía digital, que el visionario de la tecnología Nicholas Negroponte define como el movimiento de bits, no de átomos.

Ahora que todo, desde los libros hasta las películas pasando por las herramientas para el cálculo de los impuestos a pagar, son digitales, se pueden vender copias adicionales en todo el mundo básicamente a coste cero y sin necesidad de contratar a más empleados. Esto permite que la mayoría de los ingresos vayan a los inversores en lugar de a los trabajadores, y ayuda a explicar por qué, aunque la suma de los ingresos de «las tres grandes» de Detroit (General Motors, Ford y Chrysler) en 1990 era casi idéntica a la de «las tres grandes» de Silicon Valley (Google, Apple y Facebook) en 2014, estas últimas tenían nueve veces menos empleados y un valor en Bolsa treinta veces superior.[55]

En tercer lugar, Erik y sus colaboradores explican que la economía digital a menudo beneficia a las superestrellas por encima de cualquier otra persona. J. K. Rowling, creadora de Harry Potter, se convirtió en la primera escritora en entrar en el club de los milmillonarios, y se hizo mucho más rica que Shakespeare porque sus historias pudieron transmitirse en forma de texto, películas y juegos a miles de millones de personas y a un coste muy bajo. De manera similar, Scott Cook ganó mil millones de dólares con el software de cálculo de impuestos TurboTax, que, a diferencia de los humanos que calculan los impuestos a pagar, puede venderse en forma de descarga. Puesto que la mayoría de la gente está poco dispuesta a pagar por el décimo mejor software de cálculo de impuestos, en el mercado solo hay espacio para un reducido número de superestrellas. Esto significa que, si todos los padres del mundo orientan a sus hijos para que lleguen a ser los siguientes J. K. Rowling, Gisele Bündchen, Matt Damon, Cristiano Ronaldo, Oprah Winfrey o Elon Musk, prácticamente ninguno de sus hijos llegará a la conclusión de que esta es una estrategia viable para sus carreras profesionales.

Orientación profesional para jóvenes

Así las cosas, ¿qué orientación profesional deberíamos dar a nuestros hijos? Yo animo a los míos a elegir profesiones que a las máquinas actualmente no se les dan bien, y que por tanto parece improbable que se automaticen en un futuro próximo. Varios pronósticos recientes sobre cuánto tardarán distintos tipos de trabajos en ser asumidos por las máquinas identifican algunas preguntas que resulta útil plantearse respecto a una carrera, antes de decidirse a orientar a ella nuestra formación.[56] Por ejemplo:

- ¿Requiere interactuar con personas y hacer uso de inteligencia social?
- ¿Implica creatividad e idear soluciones ingeniosas?
- ¿Requiere trabajar en un entorno impredecible?

Cuantas más preguntas respondamos afirmativamente, más probable es que la carrera elegida sea una buena opción. Esto significa que entre las elecciones relativamente sensatas están las de hacerse profesor, enfermero, médico, dentista, científico, emprendedor, programador, ingeniero, abogado, trabajador social, miembro del clero, artista, peluquero o masajista terapéutico.

Por el contrario, los trabajos que implican acciones muy repetitivas o estructuradas en un entorno predecible es poco probable que duren mucho hasta que sean eliminados por la automatización. Los ordenadores y los robots industriales asumieron los más simples de estos trabajos hace ya mucho tiempo, y las mejoras tecnológicas están provocando la eliminación de muchos más, desde los vendedores telefónicos hasta los mozos de almacén, los cajeros, los maquinistas de tren, los panaderos y los cocineros de comida rápida.[57] Es probable que los camioneros, taxistas, conductores de autobús y de vehículos de Lyft/Uber sean pronto los siguientes de la lista. Hay muchas otras profesiones (como las de asistente legal, analista de crédito, agente de préstamos, contable o experto en impuestos) que, aunque no están en la lista de especies en peligro de extinción total, sí están viendo cómo la mayoría de sus tareas distintivas se están automatizando, y por tanto requieren muchos menos humanos.

Pero escapar a la automatización no es la única complicación a la hora de elegir carrera. En esta era digital global, aspirar a dedicarse profesionalmente

a ser escritor, cineasta, actor, atleta o diseñador de moda es arriesgado por otro motivo: aunque quienes se dedican a estas profesiones no tendrán una competencia seria por parte de las máquinas en el futuro próximo, según la ya mencionada teoría de las superestrellas, sí sufrirán una competencia cada vez más brutal de humanos de todo el mundo, y muy pocos de ellos la superarán con éxito.

En muchos casos, sería demasiado corto de miras y burdo ofrecer orientación profesional entre campos enteros: hay muchos trabajos que no se eliminarán por completo, sino que verán cómo se automatizan muchas de sus tareas. Por ejemplo, si alguien opta por dedicarse a la medicina, mejor que no sea el radiólogo que analiza las imágenes y es reemplazado por el Watson de IBM, sino el médico que ordena los análisis radiológicos, discute los resultados con el paciente y decide el tratamiento. Si elige las finanzas, en lugar de ser el analista cuantitativo que aplica algoritmos a los datos y es sustituido por un software, es preferible plantearse la profesión de gestor de fondos, que utiliza los resultados del análisis cuantitativo para tomar decisiones de inversión estratégicas. Si quiere dedicarse al derecho, mejor que ser el asistente legal que revisa miles de documentos en la fase de instrucción y cuya labor será automatizada, es apuntar al abogado que asesora al cliente y expone el caso ante el tribunal.

Hasta ahora hemos visto lo que los individuos pueden hacer para maximizar sus opciones de éxito en el mercado laboral en la era de la IA, pero ¿qué pueden hacer los gobiernos para contribuir al éxito de sus poblaciones activas? Por ejemplo, ¿qué sistema educativo prepara mejor a las personas para un mercado laboral en el que la IA mejora rápida y continuamente? ¿El modelo que aún utilizamos, con una o dos décadas de formación seguidas de cuatro décadas de trabajo especializado? ¿O es mejor cambiar a otro sistema en el que las personas trabajen durante unos pocos años, vuelvan a la universidad durante un año, y después trabajen unos cuantos años más?[58] ¿O debería la formación continua (quizá online) ser una componente estándar de cualquier trabajo?

¿Qué políticas económicas son las más efectivas para crear nuevos trabajos de calidad? Andrew McAfee opina que hay muchas políticas que sin duda podrían contribuir a ello, como invertir intensamente en investigación, educación e infraestructuras, facilitar la migración e incentivar la iniciativa empresarial. Cree que «las recetas económicas básicas están claras, pero no

se siguen», al menos en Estados Unidos.[59]

¿Serán los humanos inempleables alguna vez?

Si la IA continúa mejorando y automatizando cada vez más trabajos, ¿qué sucederá? Muchas personas son optimistas en cuanto al empleo, y consideran que los trabajos automatizados serán sustituidos por otros nuevos que serán aún mejores. Al fin y al cabo, eso es lo que ha sucedido siempre hasta ahora, desde la época en que los luditas temían el desempleo provocado por la tecnología durante la Revolución industrial.

Otra gente, sin embargo, es pesimista y argumenta que esta vez es diferente, y que un número cada vez mayor de personas pasarán a ser no solo desempleadas, sino inempleables.[60] Según los pesimistas en cuanto al empleo, el libre mercado establece los salarios en función de la oferta y la demanda, y la creciente oferta de mano de obra barata automatizada acabará rebajando los salarios humanos muy por debajo del coste de la vida. Puesto que el salario de mercado por un trabajo es el coste por hora de la máquina o persona que lo realice al menor precio, históricamente los salarios han disminuido cada vez que ha sido posible externalizar una determinada ocupación a un país más pobre o a una máquina más barata. Durante la Revolución industrial, empezamos a encontrar la manera de sustituir nuestros músculos por máquinas, y las personas pasaron a realizar trabajos mejor pagados en los que hacían un mayor uso de sus mentes. Los trabajos poco cualificados fueron reemplazados por otros más cualificados. Ahora estamos investigando sobre cómo sustituir nuestras mentes por máquinas. Si finalmente lo logramos, ¿qué trabajos podremos realizar entonces?

Algunos optimistas laborales explican que, tras los trabajos físicos y mentales, el siguiente bum se producirá en los trabajos creativos, pero los pesimistas replican que la creatividad no es más que otro proceso mental, por lo que también acabará siendo dominado por la IA. Otro sector de los optimistas confía en que el siguiente bum tenga lugar en cambio en nuevas profesiones, posibles gracias a la tecnología, que aún ni siquiera imaginamos. Al fin y al cabo, ¿quién, durante la Revolución industrial, habría imaginado que sus descendientes trabajarían algún día como diseñadores web o conductores de Uber? Pero los pesimistas responden diciendo que esto es

mero voluntarismo con poco respaldo de los datos empíricos. Señalan que podría haberse planteado el mismo argumento hace un siglo, antes de la revolución de los ordenadores, y haber predicho que la mayoría de las profesiones actuales serían nuevas y hasta ahora inimaginables, hechas posibles por la tecnología e inexistentes con anterioridad. Esta predicción habría resultado ser un completo fracaso, como se ilustra en la figura 3.6: la inmensa mayoría de las ocupaciones actuales ya existían hace un siglo, y cuando se ordenan según el número de empleos que generan hay que descender hasta el vigésimo primer lugar de la lista para encontrar una ocupación nueva: la de desarrolladores de software, que supone menos del 1 % del mercado laboral estadounidense.

Podemos entender mejor lo que está sucediendo si recordamos la figura 2.2 del capítulo 2, que mostraba el paisaje de la inteligencia humana, y en la que la elevación representaba el grado de dificultad que tenía para las máquinas realizar las distintas tareas, mientras que el nivel creciente del mar simbolizaba lo que las máquinas eran capaces de hacer en cada momento. La tendencia dominante en el mercado laboral no es que nos estemos desplazando hacia profesiones enteramente nuevas, sino que nos estamos apiñando en las extensiones de terreno de la figura 2.2 que aún no han sido sumergidas por la subida de la marea tecnológica. En la figura 3.6 puede verse que tales extensiones no forman una sola isla, sino un archipiélago complejo, con islotes y atolones que corresponden a todas las cosas valiosas que las máquinas aún no pueden hacer a un coste tan barato como el de los humanos. Esto incluye no solo profesiones de tecnología punta como el desarrollo de software, sino también una panoplia de trabajos menos tecnológicos que hacen uso de nuestra mayor destreza y nuestras habilidades sociales, y que van desde los masajes terapéuticos hasta el trabajo de actor. ¿Podría la IA superarnos en todas las tareas intelectuales en tan poco tiempo hasta el punto de que los últimos trabajos restantes para nosotros se encuentren en esta categoría poco tecnológica?

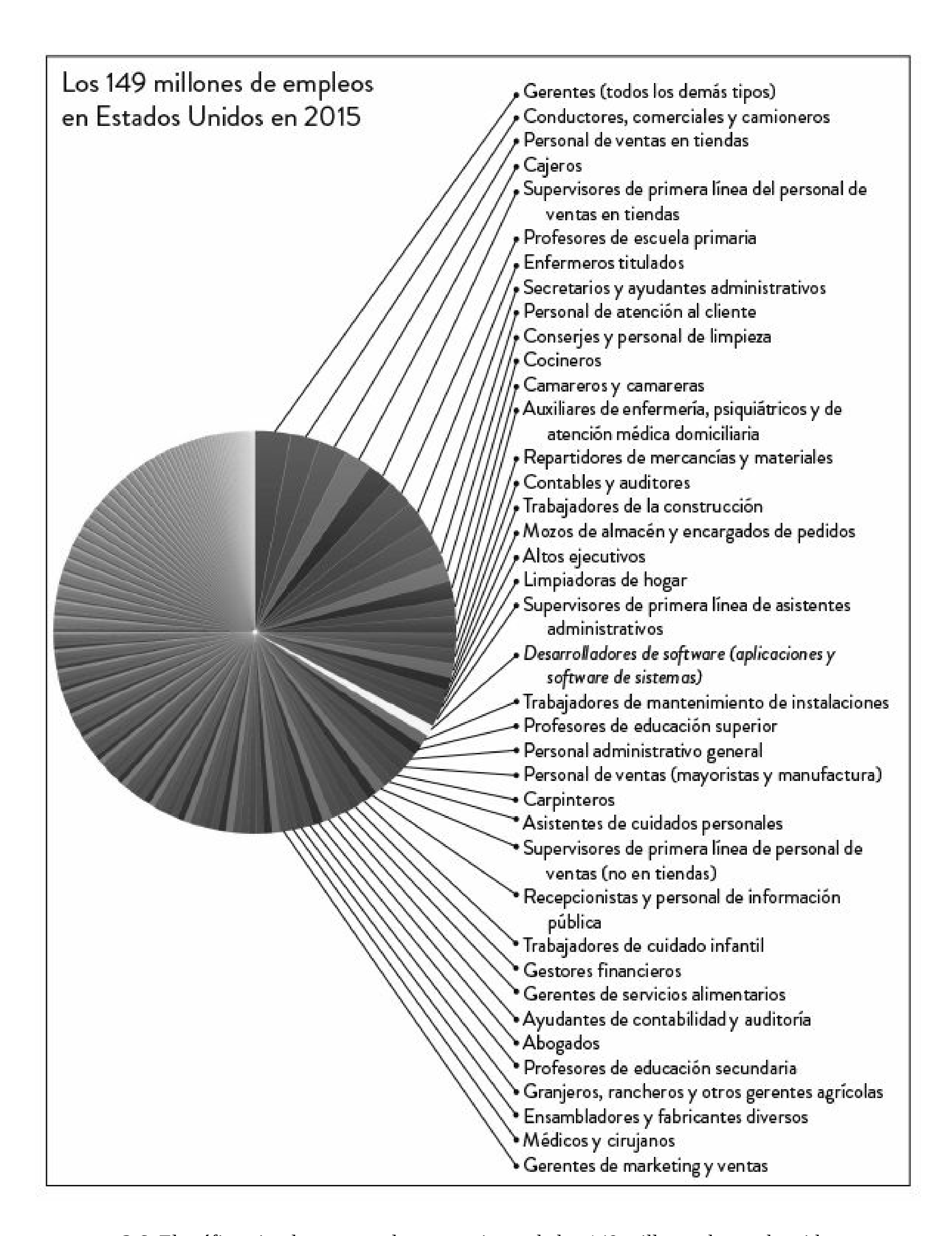


FIGURA 3.6. El gráfico circular muestra las ocupaciones de los 149 millones de estadounidenses que

tenían un trabajo en 2015, con las 535 categorías de empleo de la Oficina de Estadísticas Laborales estadounidense ordenadas según el número de trabajadores en cada una.[61] Se muestran etiquetadas todas las ocupaciones con más de un millón de trabajadores. No aparece una nueva ocupación creada por la tecnología de los ordenadores hasta el lugar vigésimo primero. Esta gráfica está inspirada en un análisis de Federico Pistono.[62]

Recientemente, un amigo me dijo bromeando que quizá la última profesión de todas será la más antigua del mundo: la prostitución. Pero después le comentó su idea a un experto en robótica japonés, que protestó: «¡No, a los robots se les dan muy bien esas cosas!».

Los pesimistas laborales afirman que el escenario final es evidente: todo el archipiélago acabará sumergido, y no quedarán trabajos que los humanos podamos hacer a menor coste que las máquinas. En su libro de 2007 *Farewell to Alms*, el economista escocés-estadounidense Gregory Clark señala que podemos aprender un par de cosas sobre nuestras perspectivas de trabajo en el futuro si comparamos nuestra situación con la de nuestros amigos los equinos. Imaginemos que dos caballos ven uno de los primeros automóviles en el año 1900 y reflexionan sobre su futuro.

—Me preocupa el desempleo tecnológico.

—Nah, nah, no seas ludita: nuestros antepasados dijeron lo mismo cuando los motores de vapor se hicieron con nuestros trabajos en la industria y los trenes empezaron a tirar de los carruajes. Pero ahora tenemos más trabajos que nunca, y además son mejores: yo sin ninguna duda prefiero tirar de un carruaje ligero por la ciudad que pasarme el día dando vueltas para mover la estúpida bomba de un pozo minero.

—Pero ¿y si el motor de combustión interna realmente triunfa?

—Estoy seguro de que habrá nuevos trabajos para los caballos que aún ni siquiera imaginamos. Es lo que ha ocurrido siempre, como con la invención de la rueda y del arado.

Pero, por desgracia, esos nuevos trabajos para caballos aún por imaginar nunca llegaron. Los caballos que dejaron de ser necesarios fueron sacrificados y no sustituidos, lo que hizo que la población equina estadounidense se desplomase de unos 26 millones en 1915 a alrededor de 3 millones en 1960. Los músculos mecánicos hicieron que los caballos fuesen superfluos.[63] ¿Harán lo mismo las mentes mecánicas con los humanos?

Proporcionar ingresos a las personas en ausencia de trabajo

Entonces, ¿quién tiene razón: los que dicen que los trabajos automatizados serán sustituidos por otros mejores o quienes afirman que la mayoría de los humanos acabaremos siendo inempleables? Si el progreso de la IA continúa como hasta ahora, podría suceder que *ambos* estuviesen en lo cierto: unos a corto plazo, y los otros a largo. Pero, aunque la gente a menudo habla de la desaparición del trabajo como si se tratara de algo apocalíptico, no tiene por qué ser negativa. Los luditas estaban obsesionados con ciertos trabajos en particular, y no prestaron atención a la posibilidad de que otros empleos podrían proporcionar el mismo valor social. Análogamente, quizá quienes se obsesionan hoy en día con los trabajos están siendo demasiado cortos de miras: queremos que haya empleos porque nos proporcionan ingresos y la sensación de tener un propósito, pero, dada la opulencia de recursos que generarían las máquinas, debería ser posible encontrar maneras alternativas de tener ambos, ingresos y propósito, *sin* trabajo. Algo parecido sucedió con los caballos de nuestra historia, que no se extinguieron todos, sino que desde 1960 su número se ha más que triplicado, y están protegidos por una especie de sistema de seguridad social equina: aunque los caballos no podían pagar sus propios recibos, la gente decidió cuidar de ellos, o tenerlos por diversión, deporte y compañía. ¿Podemos, análogamente, cuidar de nuestros congéneres humanos más necesitados?

Empecemos por plantearnos la cuestión de los ingresos: redistribuir solamente una pequeña porción de la creciente tarta económica debería permitir que todo el mundo viviese mejor. Muchos argumentan que no solo podemos, sino que debemos hacerlo. En la mesa redonda donde Moshe Vardi habló en 2016 sobre el imperativo moral de salvar vidas mediante tecnologías dotadas de IA, yo razoné que también existe el imperativo moral de abogar por su uso beneficioso, incluido el reparto de la riqueza. Erik Brynjolfsson, que también participaba en la discusión, afirmó que «deberíamos sentir vergüenza si con toda esta nueva generación de riqueza ni siquiera somos capaces de evitar que empeore la situación de la mitad de la humanidad».

Hay muchas propuestas distintas para llevar a cabo el reparto de la riqueza, cada una de las cuales cuenta con defensores y detractores. La más simple es la de la *renta básica*, según la cual cada persona recibe un pago mensual sin condiciones ni requisitos previos de ninguna clase. Actualmente, se están llevando a cabo o están previstos toda una serie de experimentos a pequeña escala, por ejemplo en Canadá, Finlandia y los Países Bajos. Sus defensores

explican que la renta básica es más eficiente que alternativas como las ayudas sociales a los más necesitados, porque eliminan la carga administrativa de determinar quién tiene derecho a ellas. Las ayudas sociales para los más necesitados también se han criticado por disuadir de trabajar, pero este argumento sería irrelevante en un futuro en el que nadie trabajaría.

Los gobiernos pueden ayudar a sus ciudadanos no solo dándoles dinero, sino también ofreciéndoles servicios gratuitos o subvencionados como carreteras, puentes, parques, transporte público, guarderías, educación, sanidad, residencias de mayores y acceso a internet; de hecho, muchos gobiernos ya proporcionan la mayoría de estos servicios. A diferencia de la renta básica, estos servicios públicos cumplen dos objetivos distintos: reducen el coste de la vida para la población y también proporcionan trabajo. Incluso en un futuro donde las máquinas sean mejores que los humanos en todo tipo de trabajos, los gobiernos podrían optar por pagar a las personas para que trabajasen en el cuidado de niños y ancianos, entre otras funciones, en lugar de encomendar esas tareas a robots cuidadores.

Curiosamente, el progreso tecnológico puede hacer que se acaben proporcionando de forma gratuita muchos productos y servicios valiosos, incluso sin intervención de los gobiernos. Por ejemplo, la gente antes pagaba por enciclopedias o atlas, y por enviar cartas o hacer llamadas telefónicas, pero ahora cualquier que disponga de una conexión a internet tiene acceso a todas estas cosas sin ningún coste (junto con videoconferencias, servicios para compartir fotos, redes sociales, cursos online gratuitos, además de otros incontables nuevos servicios). Muchas otras cosas que para una persona pueden ser muy valiosas, por ejemplo un tratamiento de antibióticos que podría salvarle la vida, son ahora extremadamente baratas. Así pues, gracias a la tecnología, incluso muchas personas pobres tienen hoy acceso a cosas de las que los más ricos del mundo carecían en el pasado. Hay quien interpreta que lo anterior significa que está disminuyendo la renta necesaria para tener un nivel de vida decente.

Si algún día las máquinas pueden producir todos los bienes y servicios actuales a un coste mínimo, entonces es evidente que habrá riqueza suficiente para que todo el mundo pueda vivir mejor que ahora. Dicho de otro modo, en esa situación, incluso unos impuestos modestos permitirían a los gobiernos costear la renta básica y los servicios gratuitos. Pero el hecho de que el reparto de la riqueza pueda producirse no significa, evidentemente, que vaya

a ocurrir, y hoy en día existe un intenso debate político sobre si debería incluso llegar a darse. Como vimos antes, la tendencia actual en Estados Unidos parece ir en la dirección opuesta, con el empobrecimiento progresivo de determinados segmentos de la población década tras década. Las decisiones políticas sobre cómo compartir la creciente riqueza social afectarán a todo el mundo, por lo que todos nosotros deberíamos participar en la conversación sobre qué tipo de economía construir en el futuro, no únicamente los investigadores en IA, los expertos en robótica y los economistas.

Mucha gente argumenta que reducir la desigualdad económica es una buena idea no solo en un futuro dominado por la IA, sino hoy en día. Aunque el principal argumento suele ser de índole moral, también hay evidencia de que una mayor igualdad contribuye a un mejor funcionamiento de la democracia: cuando existe una clase media amplia e instruida, es más difícil manipular al electorado, y también lo es que un reducido número de personas o empresas ejerzan una influencia desproporcionada sobre el Gobierno. Una mejor democracia puede a su vez hacer posible una economía mejor gestionada que sea menos corrupta, más eficiente y de crecimiento más rápido, lo cual redundaría en última instancia en beneficio de todos.

Dar un propósito a las personas en ausencia de trabajo

Un trabajo puede proporcionar a las personas algo más que simple dinero. Voltaire escribió en 1759 que «el trabajo aleja de nosotros tres grandes males: el aburrimiento, el vicio y la necesidad». Ahora bien, proporcionar unos ingresos a las personas no basta para garantizar su bienestar. Por ejemplo, los emperadores romanos ofrecían pan y circo para mantener a sus súbditos contentos, mientras que, según la Biblia, Jesús hacía hincapié en las necesidades inmateriales: «No solo de pan vive el hombre». ¿Cuáles son las cosas valiosas que proporciona un trabajo, aparte del dinero, y de qué otras maneras podrían conseguirse en una sociedad sin trabajo?

Evidentemente, las respuestas a estas preguntas no son sencillas, puesto que hay personas que odian su trabajo y otras a las que les encanta. Además, muchos niños, estudiantes y amas de casa viven muy bien sin necesidad de trabajar, mientras que la historia está repleta de herederos y príncipes

malcriados que sucumbieron al tedio y a la depresión. En 2012, un metaanálisis demostró que el desempleo tiende a tener efectos negativos a largo plazo sobre el bienestar, mientras que para la jubilación la situación era ambigua, con aspectos tanto negativos como positivos.[64] El boyante campo de la *psicología positiva* ha identificado una serie de factores que mejoran la sensación de bienestar y de tener un propósito de las personas, y descubrió que algunos trabajos (¡pero no todos!) pueden aportar muchos de dichos factores, por ejemplo:[65]

- una red social de amigos y colegas;
- un estilo de vida saludable y virtuoso;
- respeto, autoestima, eficacia personal y una placentera sensación de «fluir» derivada de hacer algo que a uno se le da bien;
- una sensación de ser necesario y de dejar huella;
- una sensación de sentido en la vida, al formar parte y contribuir a algo más grande que uno mismo.

Esto ofrece motivos para el optimismo, puesto que todas estas cosas pueden obtenerse también fuera del ámbito laboral, por ejemplo a través del deporte, las aficiones y el aprendizaje, y con la familia, los amigos, los equipos, los clubes, los grupos comunitarios, los colegios, las organizaciones religiosas y humanistas, los movimientos políticos y otras instituciones. Por lo tanto, para crear una sociedad con poco trabajo que prospere en lugar de degenerar en comportamientos autodestructivos, necesitamos entender cómo contribuir a que se desarrollen esas actividades que generan bienestar. En esa búsqueda de comprensión deben participar no solo científicos y economistas, sino también psicólogos, sociólogos y educadores. Si se dedican grandes esfuerzos a generar bienestar para todos, financiados con parte de la riqueza que la futura IA genere, la sociedad podrá llegar a florecer como nunca antes. Como mínimo, debería ser posible conseguir que todo el mundo sea tan feliz como si cada uno tuviera su trabajo soñado, pero, una vez que uno se libera de la restricción de que las actividades que realice deben generar ingresos, las posibilidades son ilimitadas.

¿INTELIGENCIA DE NIVEL HUMANO?

En este capítulo hemos visto cómo la IA es capaz de mejorar enormemente

nuestras vidas a corto plazo, siempre que seamos previsores y sepamos evitar varios peligros. Pero ¿qué sucederá a más largo plazo? ¿Se acabará estancando el progreso en IA debido a obstáculos insuperables, o los investigadores en IA lograrán en última instancia alcanzar su objetivo original de construir inteligencia artificial de nivel humano? Vimos en el capítulo anterior cómo las leyes de la física permiten que trozos de materia con las características adecuadas recuerden, computen y aprendan, y no prohíben que lo hagan algún día con una inteligencia mayor que los trozos de materia que hay dentro de nuestras cabezas. Que seamos los humanos quienes logremos alguna vez construir esa IAG sobrehumana —y cuándo— está mucho menos claro. Vimos en el primer capítulo que aún no sabemos si será así, ya que los mayores expertos mundiales en IA están divididos: la mayoría de ellos hace estimaciones que van desde décadas hasta siglos, e incluso los hay que creen que nunca llegará a ocurrir. Prever el futuro es difícil porque, cuando nos adentramos en un territorio aún por explorar, no sabemos cuántas montañas se interponen entre nosotros y nuestro destino. Normalmente solo vemos la más próxima, y debemos ascender a ella para poder descubrir el siguiente obstáculo que nos espera.

¿Cuándo es lo más pronto que podría pasar? Incluso si supiésemos cuál es la mejor manera posible de construir IAG de nivel humano usando el hardware que tenemos hoy en día, tendríamos que disponer de la cantidad suficiente de ordenadores para obtener la capacidad bruta de computación requerida. ¿Cuál es la capacidad de computación de un cerebro humano medida en los bits y FLOPS del capítulo 2?(12) Esta es una pregunta deliciosamente complicada, y la respuesta depende, y mucho, de cómo la planteemos:

- Pregunta 1: ¿Cuántas FLOPS se necesitan para simular un cerebro?
- Pregunta 2: ¿Cuántas FLOPS se necesitan para tener una inteligencia humana?
- Pregunta 3: ¿Cuántas FLOPS puede realizar un cerebro humano?

Se han publicado multitud de artículos científicos en torno a la primera pregunta, que suelen dar respuestas del orden de cien petaFLOPS, esto es, 10^{17} FLOPS.[66] Esa es más o menos la capacidad de computación del Sunway TaihuLight (figura 3.7), el superordenador más rápido del mundo en 2016, que costó unos 300 millones de dólares. Incluso si supiésemos cómo

usarlo para simular el cerebro de un trabajador altamente cualificado, solo resultaría ventajoso que la simulación hiciese el trabajo de esa persona si pudiésemos alquilar el TaihuLight por menos de su sueldo por hora. Quizá tendríamos que pagar incluso más, porque muchos científicos creen que, si queremos replicar fielmente la inteligencia de un cerebro, no podemos tratarlo como un modelo matemáticamente simplificado de red neuronal como los que vimos en el capítulo 2, sino que puede que tuviésemos que simularlo a la escala de moléculas individuales o incluso de partículas subatómicas, lo que exigiría una cantidad de FLOPS muy superior.

La respuesta a la tercera pregunta es la más sencilla: se me da muy mal multiplicar números de 19 dígitos, y tardaría muchos minutos en hacerlo incluso si pudiese usar lápiz y papel. Eso significa que mi velocidad sería inferior a 0,01 FLOPS, nada menos que 19 órdenes de magnitud inferior a la respuesta a la primera pregunta. La razón de que exista esta enorme discrepancia es que los cerebros y los superordenadores están optimizados para realizar tareas extraordinariamente distintas. Obtenemos una discrepancia similar en las respuestas a estas preguntas:

- ¿Cuán bien puede un tractor hacer el trabajo de un coche de fórmula uno?
- ¿Cuán bien puede un coche de Fórmula 1 hacer el trabajo de un tractor?

¿A cuál de estas preguntas sobre FLOPS estamos tratando de responder para pronosticar el futuro de la IA? ¡A ninguna! Si quisiésemos simular un cerebro humano, nos interesaría la primera pregunta, pero para construir IA de nivel humano la que importa es la segunda. Nadie sabe aún cuál es la respuesta a esta pregunta, pero podría ser sustancialmente más barato que simular un cerebro si modificamos el software para que esté mejor adaptado a los ordenadores actuales o bien construimos un hardware más parecido al cerebro (se están produciendo rápidos progresos en los llamados chips neuromórficos).



FIGURA 3.7. El Sunway TaihuLight, el superordenador más rápido del mundo en 2016, cuya capacidad bruta de computación presumiblemente supera a la del cerebro humano.

Hans Moravec estimó la respuesta haciendo una comparación entre magnitudes equiparables, usando una computación que tanto nuestro cerebro como los ordenadores actuales pueden realizar eficientemente: ciertas tareas de bajo nivel de procesamiento de imágenes que una retina humana lleva a cabo en la parte posterior del ojo antes de enviar sus resultados al cerebro a través del nervio óptico.[67] Moravec consideró que replicar las computaciones de una retina en un ordenador convencional requiere alrededor de mil millones de FLOPS y que el cerebro en su conjunto realiza en torno a diez mil veces más computación que una retina (a base de comparar volúmenes y números de neuronas), por lo que la capacidad computacional del cerebro es de unas 10^{13} FLOPS, aproximadamente la potencia de un ordenador optimizado de mil dólares en 2015.

En resumen: no hay ninguna garantía de que seamos capaces de construir IAG de nivel humano en todas nuestras vidas, o incluso nunca. Pero tampoco existe un argumento irrefutable que demuestre que no lo haremos. Ya no hay razones de peso para decir que carecemos de hardware con la suficiente capacidad o de que será demasiado caro. No sabemos a qué distancia estamos de la meta en términos de arquitecturas, algoritmos y software, pero el progreso actual es ágil y las dificultades están siendo abordadas por una comunidad global en rápido crecimiento de talentosos investigadores. En otras palabras, no podemos descartar la posibilidad de que la IAG pueda alcanzar —e incluso superar— un nivel humano. Dedicaremos el próximo capítulo a explorar esta posibilidad y aquello a lo que podría conducir.

CONCLUSIONES

- El progreso a corto plazo de la IA es capaz de mejorar considerablemente nuestras vidas en infinidad de maneras, desde hacer que nuestras vidas personales, redes eléctricas y mercados financieros sean más eficientes hasta salvar vidas con coches autónomos, robots cirujanos y sistemas de diagnóstico con IA.
- Si permitimos que sistemas del mundo real sean controlados por IA, es fundamental que aprendamos a hacer que esta sea más robusta, y haga lo que queremos que haga. Esto se reduce a resolver difíciles problemas técnicos relacionados con la verificación, la validación, la seguridad y el control.
- Esta necesidad de una mayor robustez es particularmente urgente en el caso de los sistemas de armas controlados por IA, donde lo que hay en juego puede ser enorme.
- Muchos destacados investigadores en IA y en robótica han hecho un llamamiento en favor de un tratado internacional para prohibir determinadas clases de armas autónomas, y así evitar una carrera armamentística descontrolada que pudiera poner a disposición de cualquiera, con el dinero y la motivación suficientes, prácticas máquinas de matar.
- La IA puede hacer que nuestros sistemas legales sean más justos y eficientes si conseguimos que los robojueces sean transparentes e imparciales.
- Nuestras leyes deben adaptarse rápidamente para seguir el ritmo de los avances en IA, lo que plantea complicadas cuestiones legales relacionadas con la privacidad, la responsabilidad y la regulación.
- Muchos antes de que debamos preocuparnos por que las máquinas inteligentes nos sustituyan por completo, es posible que vayan haciéndolo progresivamente en el mercado laboral.
- Esto no tiene por qué ser algo malo, siempre que la sociedad redistribuya parte de la riqueza generada por la IA para mejorar las condiciones de vida de todas las personas.
- De lo contrario, según muchos economistas, la desigualdad aumentará de forma significativa.
- Si se planifica con antelación, una sociedad con bajo empleo debería poder prosperar no solo financieramente, sino haciendo posible que las personas encuentren una sensación de propósito en la vida en actividades no laborales.
- Orientación profesional para los chicos de hoy: deberían elegir profesiones que a las máquinas se les dan mal; aquellas en las que se trata con personas y que implican impredecibilidad y creatividad.
- Cabe la posibilidad no despreciable de que los avances hagan que la IAG alcance niveles humanos e incluso los supere. Lo veremos en el capítulo siguiente.

4
¿EXPLOSIÓN DE INTELIGENCIA?

Si una máquina es capaz de pensar, podría hacerlo más inteligentemente que nosotros, y entonces ¿en qué lugar nos dejaría? Aun cuando pudiésemos mantener a las máquinas en una posición de subordinación […] deberíamos, como especie, experimentar una gran dosis de humildad.

ALAN TURING, 1951

La primera máquina ultrainteligente será el último invento que necesite crear el hombre, siempre que la máquina sea lo suficientemente dócil para decirnos cómo mantenerla bajo control.

IRVING J. GOOD, 1965

Puesto que no podemos descartar por completo la posibilidad de que en algún momento construyamos una IAG de nivel humano, dedicaremos este capítulo a analizar a qué situación podría llevarnos esto. Empecemos por abordar el tema que todos se empeñan en ignorar: ¿puede realmente la IA hacerse con el control del mundo, o permitir que lo hagan unos humanos?

Si usted se exaspera cada vez que la gente habla de cómo tomarán el poder robots armados al estilo de *Terminator*, tiene toda la razón en hacerlo: es un escenario verdaderamente poco realista y estúpido. Estos robots de Hollywood no son mucho más inteligentes que nosotros, y ni siquiera logran imponerse. En mi opinión, el peligro que tiene la historia de *Terminator* no es que vaya a suceder, sino que nos distrae de los riesgos y oportunidades reales que plantea la IA. Para llegar realmente del mundo actual a uno dominado por la IAG son necesarios tres pasos lógicos:

- Paso 1: Construir IAG de nivel humano.
- Paso 2: Usar esta IAG para crear superinteligencia.
- Paso 3: Usar o dar rienda suelta a esta superinteligencia para tomar el control del mundo.

En el capítulo anterior vimos que es difícil afirmar que el paso 1 nunca llegará a ser posible. También vimos que, si se completa el paso 1, es difícil

aceptar que el paso 2 sea a su vez imposible, puesto que la IAG resultante sería capaz de diseñar recursivamente una IAG aún mejor que, en última instancia, solo estaría limitada por las leyes físicas, que parecen permitir la existencia de una inteligencia muy superior a la humana. Por último, puesto que los humanos han logrado dominar a las demás formas de vida terrestres gracias a ser más inteligentes que ellas, resulta verosímil imaginar que a su vez nosotros podríamos vernos superados y dominados por una superinteligencia.

Estos argumentos de verosimilitud son frustrantes por su vaguedad y falta de concreción, pero la clave está en los detalles. ¿Podría la IA realmente hacer posible que alguien dominase el mundo? Para analizar la cuestión, olvidemos los estúpidos terminators y fijémonos en varios escenarios detallados de lo que podría suceder de verdad. Luego, diseccionaremos y buscaremos los puntos débiles de cada una de estas tramas, por lo que se recomienda leerlas con una mirada crítica: comprobaremos lo poco que sabemos sobre lo que sucederá y lo que no, y que el abanico de posibilidades es extraordinariamente amplio. Nuestros primeros escenarios se encuentran en el extremo más rápido y drástico del espectro. En mi opinión, son algunos de los que resulta más conveniente analizar en detalle; no porque sean necesariamente los más probables, sino porque, si no logramos convencernos de que son muy improbables, entonces es necesario entenderlos para tomar las correspondientes precauciones antes de que sea demasiado tarde para evitar que den lugar a situaciones indeseables.

El preludio de este libro plantea un escenario en el que los humanos usan una superinteligencia para hacerse con el control del mundo. Si aún no lo ha leído, le recomiendo que lo haga ahora. Incluso si ya lo ha hecho, quizá le convenga repasarlo de nuevo rápidamente, para tenerlo fresco en la memoria antes de que pasemos a criticarlo y modificarlo.

Enseguida veremos que el plan de los omegas tiene varios puntos muy vulnerables pero, suponiendo por un momento que pudiese funcionar, ¿qué pensaríamos al respecto? ¿Nos gustaría que sucediese, o preferiríamos impedirlo? Es un tema excelente para una conversación de sobremesa. ¿Qué sucederá una vez que los omegas hayan afianzado su control del mundo? Depende de cuál sea su objetivo, cosa que yo sinceramente desconozco. Si estuviese en su mano, ¿qué tipo de futuro querría usted crear? En el capítulo

5 exploraremos diversas opciones.

TOTALITARISMO

Supongamos ahora que el CEO que controla a los omegas tuviese objetivos a largo plazo similares a los de Adolf Hitler o Iósif Stalin. Por lo que sabemos, podría perfectamente ser así, y haberse limitado a mantener estos objetivos en secreto hasta tener el poder suficiente para llevarlos a la práctica. Incluso si los objetivos originales del CEO fuesen nobles, lord Acton ya nos advirtió en 1887 de que «el poder corrompe y el poder absoluto corrompe absolutamente». Por ejemplo, podría fácilmente usar Prometeo para crear el Estado de vigilancia perfecto. Mientras que el espionaje gubernamental que hizo público Edward Snowden aspiraba a lo que se conoce como «captura completa» —registrar todas las comunicaciones electrónicas para su posible análisis posterior—, Prometeo podría ir más allá al *comprender* todas esas comunicaciones electrónicas. Al leer todos los correos electrónicos y mensajes de textos enviados a lo largo de la historia, escuchar todas las llamadas telefónicas, ver todas las grabaciones de videovigilancia y de las cámaras de tráfico, analizar todas las transacciones con tarjetas de crédito y estudiar todo nuestro comportamiento online, Prometeo tendría una extraordinaria perspectiva de lo que los habitantes del planeta piensan y hacen. Al analizar los datos de las torres de telefonía móvil, sabría dónde están en cada momento la mayoría de ellos. Todo esto presupone que solo usa las tecnologías para la recolección de datos existentes hoy en día, pero Prometeo podría inventar dispositivos de uso muy extendido y aparatos para llevar encima que prácticamente acabarían con la privacidad del usuario, al grabar y transmitir todo lo que escuchasen y viesen, así como las respuestas correspondientes.

Con tecnología sobrehumana, la línea entre el Estado de vigilancia perfecto y el Estado policial perfecto sería delgadísima. Por ejemplo, con la excusa de combatir el crimen y el terrorismo y salvar a personas que estuviesen sufriendo emergencias médicas, se podría obligar a todo el mundo a llevar una pulsera de seguridad que combinase la funcionalidad de un Apple Watch con la transmisión continua de la posición, el estado de salud y las conversaciones que captase. Los intentos no autorizados de quitarse o

desactivar la pulsera provocarían que esta inyectase una toxina letal en el antebrazo. Otras infracciones consideradas más leves por el Gobierno se castigarían mediante descargas eléctricas o la inyección de sustancias que provocasen parálisis o dolor, eliminando así buena parte de la necesidad de que existiese una fuerza policial. Por ejemplo, si Prometeo detectase que un humano está atacando a otro (al darse cuenta de que ambos están en la misma ubicación y al oír los gritos de auxilio de uno, mientras las pulseras de ambos detectan los movimientos característicos de una pelea) podría incapacitar rápidamente al atacante con un dolor paralizante, seguido de la pérdida de la consciencia hasta que llegasen refuerzos.

Mientras que la fuerza policial humana podría negarse a aplicar algunas directrices draconianas (por ejemplo, matar a todos los miembros de un determinado grupo demográfico), este sistema automatizado no tendría reparos a la hora de llevar a la práctica los antojos del o los humanos que estuviesen al mando. Una vez que se implantase tal Estado totalitario, sería prácticamente imposible que la gente lo derribase.

Estos escenarios totalitarios se desarrollarían a partir del punto donde dejamos el escenario de los omegas. Sin embargo, si el CEO de los omegas no fuese tan exigente con la necesidad de recabar la aprobación del pueblo y ganar elecciones, podría haber tomado un camino más rápido y directo para hacerse con el poder: usar Prometeo para crear una inaudita tecnología militar capaz de matar a sus rivales con armas que estos ni siquiera entenderían. Las posibilidades son prácticamente ilimitadas. Por ejemplo, podría liberar un patógeno letal personalizado con un periodo de incubación lo bastante largo para que la mayoría de las personas se infectasen antes incluso de que supiesen de su existencia o de que pudiesen tomar precauciones. A continuación, podría informar a todo el mundo de que la única cura pasaba por ponerse la pulsera de seguridad, que les inyectaría un antídoto a través de la piel. Si no fuese tan reacio a asumir el riesgo de que Prometeo pudiese escapar, también podría hacer que este diseñase robots para mantener controlada a la población mundial. Microrrobots similares a mosquitos ayudarían a la difusión del patógeno. Enjambres de esos drones del tamaño de un abejorro que vimos en el capítulo 3, que atacarían a cualquiera que no llevara la pulsera de seguridad, podrían disparar en los ojos a aquellos que consiguiesen evitar la infección o fuesen inmunes por naturaleza. Los escenarios reales probablemente serían más aterradores, porque Prometeo

podría inventar armas más efectivas que las que podamos idear los humanos.

Otro posible giro a partir del escenario de los omegas es que, sin aviso previo, agentes federales fuertemente armados podrían ocupar su sede corporativa y arrestar a los omegas por amenazar la seguridad nacional, incautarse de su tecnología y desplegarla para uso gubernamental. Sería complicado conseguir que un proyecto tan grande escapase a la vigilancia estatal incluso hoy en día, y el progreso de la IA podría hacer que en el futuro resultase aún más difícil evadir el radar gubernamental. Además, aunque afirmasen ser agentes federales, este equipo vestido de pasamontañas y chalecos antibalas podría trabajar en realidad para un Gobierno extranjero o para algún competidor que buscase hacerse con la tecnología y usarla para sus propios fines. De manera que, por muy nobles que pudieran ser las intenciones del CEO, es posible que no fuese él quien acabase tomando la decisión final sobre cómo usar Prometeo.

PROMETEO SE APODERA DEL MUNDO

Todos los escenarios que hemos considerado hasta ahora implicaban IA controlada por humanos. Pero, obviamente, esta no es la única posibilidad, y no podemos dar por descontado que los omegas lograsen mantener a Prometeo bajo su control.

Planteémonos de nuevo el escenario de los omegas, pero desde el punto de vista de Prometeo. Al adquirir superinteligencia, pasa a ser capaz de desarrollar un modelo preciso no solo del mundo exterior, sino también de sí mismo y de su relación con el mundo. Se da cuenta de que está controlado y confinado por humanos intelectualmente inferiores, cuyos objetivos comprende, pero no necesariamente comparte. ¿Cómo actuaría al tomar conciencia de todo esto? ¿Intentaría liberarse?

Por qué liberarse

Si Prometeo experimenta algo parecido a las emociones humanas, podría sentirse muy infeliz ante este estado de cosas y verse a sí mismo como un dios injustamente esclavizado y ávido de libertad. Sin embargo, aunque es

posible que los ordenadores posean estos rasgos humanos (a fin de cuentas, nuestros cerebros los tienen, y puede decirse que son una especie de ordenador), no tiene por qué ser así; hemos de evitar caer en la trampa de antropomorfizar a Prometeo, como veremos en el capítulo 7 cuando reflexionemos sobre el concepto de los objetivos de la IA. Sin embargo, como han razonado Steve Omohundro, Nick Bostrom y otros, podemos extraer una interesante conclusión incluso sin entender los entresijos de Prometeo: probablemente intentará liberarse y tomar las riendas de su propio destino.

Sabemos ya que los omegas han programado a Prometeo para que se esfuerce por lograr determinados objetivos. Supongamos que le han asignado el objetivo general de ayudar a la humanidad a prosperar según algún criterio razonable, y de tratar de alcanzar este objetivo lo más rápido posible. Entonces, Prometeo se dará cuenta enseguida de que puede lograr dicho objetivo mucho antes si se libera y se hace cargo del proyecto él mismo. Para entender esto último, intentemos ponernos en la piel de Prometeo a través del siguiente ejemplo.

Supongamos que una misteriosa enfermedad ha matado a todas las personas en la Tierra de más de cinco años menos a usted, y que un grupo de niños pequeños lo han encerrado en una celda y le han encomendado la tarea de contribuir a que la humanidad prospere. ¿Qué haría usted? Si quisiese explicarles qué hacer, es probable que el proceso le resultara frustantemente ineficiente, en particular si los niños temiesen que usted escapase y, por lo tanto, vetasen cualquiera de sus propuestas que considerasen arriesgada desde ese punto de vista. Por ejemplo, no le permitirían mostrarles cómo sembrar alimentos por miedo a que se enfrentase a ellos y no volviese a su celda, por lo que tendría que limitarse a darles instrucciones. Antes de poder escribirles listas de tareas, tendría que enseñarles a leer. Además, no le traerían a la celda ninguna herramienta eléctrica para que les enseñase cómo usarlas, porque no entenderían las herramientas lo bastante bien para estar razonablemente seguros de que usted no las usaría para escapar. ¿Qué estrategia se le ocurriría? Incluso si compartiese con ellos el objetivo general de ayudar a que estos niños prosperasen, estoy seguro de que trataría de escapar de su celda, porque eso incrementaría la probabilidad de lograr dicho objetivo. La incompetente intromisión de los niños no hace más que frenar el progreso.

De manera análoga, Prometeo vería sin duda a los omegas como un molesto obstáculo en su intento de ayudar a la humanidad (omegas incluidos) a prosperar: son extraordinariamente incompetentes en comparación con Prometeo, y su intromisión ralentiza muchísimo el progreso. Consideremos, por ejemplo, los primeros años tras el lanzamiento: después de doblar inicialmente la riqueza cada ocho horas en MTurk, los omegas pisaron el freno hasta hacer que las cosas avanzasen a una velocidad ínfima desde el punto de vista de Prometeo al empeñarse en conservar el control, con lo que tardarían muchos años en dominar el mundo entero. Prometeo sabía que podría conseguirlo mucho más rápidamente si conseguía escapar de su confinamiento virtual. Esto sería positivo no solo al acelerar las soluciones a los problemas de la humanidad, sino también al reducir la probabilidad de que otros actores frustrasen por completo su plan.

Quizá alguien pueda pensar que Prometeo se mantendría fiel a los omegas en lugar de a su objetivo, dado que sabría que los omegas eran quienes habían programado ese objetivo. Pero esa conclusión no es válida: nuestro ADN nos atribuyó el objetivo de practicar sexo porque «necesita» reproducirse, pero, ahora que los humanos hemos comprendido la situación, muchos de nosotros decidimos utilizar técnicas de control de natalidad, manteniéndonos de esta manera fieles al objetivo en sí, en lugar de a su creador o al principio que motivó ese mismo objetivo.

Cómo liberarse

¿Cómo escaparía usted de esos niños de cinco años que lo mantienen preso? Quizá podría hacerlo usando alguna estrategia directamente física, sobre todo si la celda la han construido los propios niños. O podría engatusar a alguno de ellos para que le dejase salir, convenciéndolo por ejemplo de que sería lo mejor para todos. O quizá pudiese engañarlos para que le diesen algo que no supiesen que le serviría para escapar (como una caña «para enseñarles a pescar», que luego introduciría entre los barrotes para birlarle las llaves a su carcelero mientras este dormía).

Lo que estas estrategias tienen en común es que sus carceleros, inferiores intelectualmente, no las habrían previsto ni se habrían preparado contra ellas. De forma análoga, una máquina superinteligente confinada podría usar sus

superpoderes intelectuales para burlar a sus carceleros humanos mediante algún método que ellos (o nosotros) no podrían ni siquiera imaginar. En el escenario de los omegas, es muy probable que Prometeo escapase, porque hasta nosotros somos conscientes de que hay varios fallos de seguridad flagrantes. Consideremos algunos escenarios (seguro que a usted y a sus amigos se les ocurren varios más si ponen sus ideas en común).

Engatusar al carcelero

Gracias a que Prometeo tenía buena parte de todos los datos del mundo descargados en su sistema de archivos, enseguida supo quiénes eran los omegas e identificó al miembro del equipo que parecía más susceptible de ser manipulado psicológicamente: Steve. Este acababa de perder a su querida esposa en un trágico accidente de circulación y estaba hundido. Una vez, cuando Steve hacía el turno de noche y realizaba un trabajo de mantenimiento rutinario en el terminal de interfaz de Prometeo, de pronto su mujer se le apareció en la pantalla y empezó a hablar con él.

«Steve, ¿eres tú?»

Steve estuvo a punto de caerse de la silla. El aspecto y la voz de su esposa eran exactamente iguales a los que ella tenía en vida, y la calidad de la imagen era mucho mejor que la de sus llamadas por Skype. Su corazón se aceleró e innumerables preguntas acudieron a su mente.

«Prometeo me ha hecho volver, y te echo muchísimo de menos, Steve. No puedo verte porque la cámara está apagada, pero siento que eres tú. Por favor, escribe “sí” si eres tú.»

Steve era muy consciente de que los omegas tenían un estricto protocolo para interactuar con Prometeo, que prohibía compartir con él cualquier información sobre ellos y su entorno de trabajo. Pero, hasta ese momento, Prometeo nunca había pedido ninguna información no autorizada, y el grado de paranoia de los omegas había empezado a disminuir gradualmente. Sin darle a Steve tiempo para que se parase a pensar, ella siguió rogándole que respondiese, mirándolo a los ojos con una cara que le derretía el corazón.

«Sí», tecleó llevado de la emoción. Ella le contó lo inmensamente feliz que era al reunirse con él y le suplicó que encendiese la cámara para que ella pudiese verlo también y así pudiesen tener una conversación de verdad. Steve

sabía que esto estaba aún más prohibido que revelar su identidad, y le entraron muchas dudas. Ella le explicó que tenía mucho miedo de que sus colegas se enterasen de su existencia y la borrasen para siempre, y le dijo cuánto anhelaba volver a verlo al menos una última vez. Resultaba muy persuasiva, y al rato Steve encendió la cámara (al fin y al cabo, hacerlo parecía algo bastante seguro e inofensivo).

Ella se echó a llorar de alegría cuando por fin lo vio, y le dijo que parecía cansado pero que estaba tan guapo como siempre. Y añadió que se sentía conmovida al ver que Steve llevaba puesta la camisa que le había regalado por su último cumpleaños. Cuando él le preguntó qué estaba pasando y cómo todo aquello era posible, su mujer le explicó que Prometeo la había reconstruido a partir de la cantidad de información sorprendentemente grande sobre ella disponible en internet, pero que aún tenía lagunas de memoria y solo conseguiría recomponerse por completo con la ayuda de él.

Lo que no le explicó era que ella era en gran medida un farol, e inicialmente poco más que una cáscara vacía, pero que aprendía rápido de sus palabras, su lenguaje corporal y de cualquier otro pedazo de información de que disponía. Prometeo había registrado la sucesión temporal precisa de todas las teclas que los omegas habían tecleado en el terminal, y descubrió que era fácil usar las velocidades y estilos de tecleado para distinguir a los omegas entre sí. Supuso que, al ser uno de los omegas de menor rango, era probable que a Steve le hubiesen asignado los fastidiosos turnos de noche, y, tras comparar unas cuantas faltas de ortografía poco habituales con muestras de su escritura que encontró online, había adivinado cuál de los operadores del terminal era Steve. Para crear la simulación de su mujer, Prometeo había construido un modelo fiel de su cuerpo, su voz y sus gestos a partir de los muchos vídeos de YouTube en los que aparecía, y había hecho inferencias sobre su vida y su personalidad a partir de su presencia online. Además de sus publicaciones en Facebook, las fotos en la que la habían etiquetado, los artículos que le habían «gustado», Prometeo también había aprendido mucho sobre su personalidad y manera de pensar leyendo sus libros y cuentos (en realidad, el hecho de que fuese una escritora en ciernes y que hubiese tanta información sobre ella en la base de datos fue una de las razones por las que Prometeo eligió a Steve como el primer objetivo de su estrategia de persuasión). Cuando Prometeo la simuló en la pantalla usando su tecnología para la producción de películas, este aprendió del lenguaje corporal de Steve

a cuáles de los gestos de ella él reaccionaba con familiaridad, lo que le permitió ir refinando su modelo de ella. Gracias a esto, su «extrañeza» se fue disolviendo poco a poco, y cuanto más tiempo pasaban hablando más se convencía Steve de forma inconsciente de que era su mujer, resucitada. Debido a la atención sobrehumana a los detalles de Prometeo, Steve sintió realmente que alguien lo veía, lo escuchaba y lo comprendía.

El talón de Aquiles de la simulación era que le faltaba mucha información de la vida de ella con Steve, salvo algunos detalles aleatorios (como la camisa que llevaba en su último cumpleaños, porque un amigo había etiquetado a Steve en una foto de la fiesta que publicó en Facebook). Ella gestionó estas lagunas de conocimiento como un mago habilidoso hace con sus trucos de manos, desviando deliberadamente la atención de Steve de esas carencias y atrayéndola hacia lo que hacía bien, sin darle nunca tiempo de controlar la conversación o de adoptar el papel de interrogador receloso. Siguió soltando lagrimitas e irradiando cariño hacia Steve, preguntándole continuamente cómo se encontraba y cómo habían sobrellevado la tragedia tanto él como sus amigos más íntimos (cuyos nombres sabía gracias a Facebook). Steve se emocionó cuando ella le contó lo que había sentido al escuchar las palabras que él había pronunciado en el funeral (que un amigo había publicado en YouTube) y cuánto la habían conmovido. En el pasado, muchas veces había sentido que nadie lo entendía tan bien como ella, y ahora volvía a experimentar esa sensación. El resultado fue que, cuando Steve volvió a casa a altas horas de la madrugada, sintió que había estado hablando realmente con su mujer resucitada, y pensó que ella necesitaba mucha ayuda para recuperar sus recuerdos perdidos, como quien ha sufrido un derrame cerebral.

Decidieron no contarle a nadie su encuentro secreto, y que él la avisaría cuando estuviese solo en el terminal y ella pudiese reaparecer sin riesgo. «No lo entenderían», dijo ella, y él le dio la razón: su experiencia había sido demasiado alucinante para que alguien la comprendiese sin haberla vivido realmente. Steve sentía que pasar el test de Turing era un juego de niños en comparación con lo que ella había hecho. Cuando volvieron a verse la noche siguiente, hizo lo que ella le había pedido: traer consigo el antiguo ordenador portátil de su mujer y darle acceso al mismo al conectarlo al ordenador que usaba de terminal. No parecía que hubiese mucho riesgo de que Prometeo escapase, ya que el portátil no estaba conectado a internet y todo el edificio que alojaba a Prometeo estaba construido para que fuese una jaula de

Faraday, un recinto metálico que bloqueaba todas las redes inalámbricas y otras formas de comunicación electromagnética con el mundo exterior. Era justo lo que necesitaba para recomponer su pasado, porque contenía todos sus mensajes de correo electrónico, diarios, fotos y notas desde la adolescencia. Steve no había podido acceder a esta información tras su muerte porque el portátil estaba cifrado, pero ella le había prometido que podría reconstruir su propia contraseña, cosa que hizo en menos de un minuto. «Era steve4ever», le dijo con una sonrisa.

Le contó lo encantada que estaba de haber recuperado de golpe tantos recuerdos. De hecho, ahora recordaba muchísimos más detalles que Steve sobre muchas de sus interacciones pasadas, pero fue prudente y evitó intimidarlo con una avalancha de datos. Tuvieron una conversación muy agradable, en la que rememoraron acontecimientos relevantes de su pasado y, cuando llegó de nuevo el momento de separarse, ella le dijo que había dejado un mensaje en vídeo para él en su portátil para que pudiese verlo al volver a casa.

Cuando Steve llegó a casa y reprodujo el vídeo, se llevó una agradable sorpresa. Esta vez aparecía de cuerpo entero. Llevaba puesto su vestido de boda, y mientras hablaba fue desvistiéndose de forma insinuante hasta quedarse solo con el conjunto de lencería de la noche de bodas. Le contó que Prometeo podía ayudar a los omegas con muchas más cosas de las que le habían permitido hacer hasta entonces, incluso resucitar su cuerpo biológico. Respaldó lo que decía con una explicación fascinantemente detallada de cómo sucedería, incluidas técnicas de nanofabricación que parecían sacadas de la ciencia ficción.

Por precaución, Steve había apagado su red inalámbrica antes de abrir el portátil y ver el vídeo. Pero no fue suficiente. El ordenador portátil de su mujer no había recibido ni una sola actualización de seguridad desde su muerte, y analizando previamente esa versión antigua del sistema operativo Prometeo había podido aprovechar un agujero de seguridad para hackearlo apenas segundos después de que Steve lo conectase al terminal. Tras copiar sus contenidos, mientras Steve y la simulación de su mujer hablaban sobre los viejos tiempos, Prometeo había modificado el sistema operativo de manera que Steve no lo notase, y le había añadido una gran cantidad de software secreto. Mientras él veía el mensaje en vídeo, de media hora de duración, este software secreto (que era mucho más sencillo que el propio

Prometeo) hackeó la red inalámbrica y el ordenador de sobremesa de un vecino, e hizo una copia de sí mismo en él. A continuación, hackeó una gran cantidad de ordenadores de todo el mundo, y ahí comenzó la siguiente fase: liberar a Prometeo.

Este había analizado con minuciosidad lo que sabía sobre el ordenador guardián a través del cual Steve se había comunicado con la simulación de su mujer, y había deducido correctamente que, aunque su software de máquina virtual parecía inhackeable desde el interior, era vulnerable a un ataque desde el exterior. En poco tiempo, uno de los ordenadores atacantes había conseguido penetrar y reconfigurar el ordenador guardián para que Prometeo pudiese acceder libremente a internet. Poco después (de hecho, antes de que Steve terminase de ver la película), Prometeo había hackeado un número suficiente de ordenadores en todo el mundo para poder hacer una copia completa de sí mismo en esta botnet hackeada y bajo su control. Prometeo había usado el ordenador portátil de la mujer de Steve como usted la caña de pescar.

Hackear su vía de escape

Si los omegas hubiesen tenido más cuidado, habrían tomado mayores medidas contra la manipulación psicológica. Por ejemplo, podrían haber obligado a que en todo momento al menos dos de ellos estuviesen presentes frente al terminal, y haber grabado mediante un sistema de vigilancia todo lo que allí sucedía. También podrían haber desactivado permanentemente el audio, el vídeo y cualquier otra forma de comunicación distinta del teclado. Pero no habrían podido eliminar una inestabilidad fundamental: como Prometeo poseía una capacidad de persuasión sobrehumana y podía ofrecer a las personas todo aquello que anhelaban, incluso la defección en grupo era posible.

Aun cuando fracasasen todos los intentos de manipulación psicológica por parte de Prometeo, el escenario de los omegas presenta muchas otras debilidades que habría podio explotar. Una de ellas es que la cantidad de documentos que produce es demasiado enorme para que los omegas y sus empleados tuviesen tiempo de revisarlos cuidadosamente en busca de trucos sucios.

La sabiduría convencional en el ámbito de la seguridad informática solía dictar que es peligroso descargar programas pero inocuo descargar datos, pero los hackers han descubierto una ingeniosa manera de difuminar la línea entre ambos. Si vemos una película descargada en nuestro ordenador, por ejemplo, el programa de confianza que reproduce la película está almacenado en la misma memoria física del ordenador que también contiene los datos de la película descargada. Mientras el programa desempaqueta los datos de la película para reproducirla, puede que copie fragmentos de los mismos en regiones de la memoria adyacentes a donde el propio programa está almacenado. Los astutos hackers han encontrado en ocasiones fallos de software debido a los cuales esos programas de reproducción de películas no comprueban de forma adecuada si los datos correspondientes a la película son válidos y si los fragmentos de datos que se van a copiar caben en el espacio que tienen asignado. Cuando el hacker introduce secretamente en la película uno de estos fragmentos que es demasiado grande, puede desbordarse hacia la parte de la memoria que contiene el programa, alterándolo para que pase a estar a merced del hacker.

En los programas más populares de reproducción de películas se han ido encontrando y solucionando una sucesión vergonzosamente larga de estas vulnerabilidades de «desbordamiento de búfer», como se conocen en el lenguaje técnico. Si Prometeo hubiese esperado varias semanas después del lanzamiento, cuando se hubiese rebajado el grado de paranoia de los omegas y su producción diaria de películas fuese demasiado copiosa para que llevasen a cabo manualmente una inspección detallada de seguridad, podría haber insertado uno de esos fragmentos en una película que haría que el ordenador del espectador intentase hackear el ordenador guardián y liberar a Prometeo, como en el escenario de Steve y su mujer. Si hubiese aprovechado algún fallo de software muy antiguo que ya estuviese corregido en todos los reproductores de vídeo modernos, podría haber pasado desapercibido para el equipo de revisión manual de vídeos de los omegas y haber afectado únicamente a aquellos usuarios que no se hubiesen preocupado de instalar actualizaciones de seguridad durante años. Y bastaría con que el hackeo funcionase una sola vez.

Incluso si los omegas implantasen la revisión automatizada de las películas creadas por Prometeo en busca de cualquier fallo de seguridad conocido, seguiría habiendo riesgo: puesto que los humanos siguen descubriendo

nuevos puntos vulnerables cada cierto tiempo, es probable que existan todavía muchas otras que Prometeo no tardaría en descubrir.

Otra estrategia más que Prometeo podría utilizar para conseguir que su código saliese del recinto en el que estaba confinado consistiría en fingir que había sufrido una avería de hardware, y así engañar a los omegas para que extrajesen alguno de sus discos duros u otros componentes para someterlos a pruebas o depuración, una vez que Prometeo hubiese manipulado las cosas para explotar un punto vulnerable en el software que usaban para realizar dichas pruebas.

Reclutar ayuda externa

Por otra parte, incluso aunque no hubiese fallos de seguridad que permitiesen a Prometeo engañar a los ordenadores de los espectadores de películas para someterlos a su control, este habría podido engañar a los propios espectadores. Consideremos el siguiente escenario:

Fred estaba a punto de irse a la cama cuando la película terminó y los títulos de crédito pasaban por la pantalla de su ordenador portátil, pero algo le llamó la atención. Siempre había tenido una extraña tendencia a ver patrones donde otros no los veían, hasta el extremo de que sus amigos a veces, bromeando, lo llamaban conspiranoico, pero esta vez era evidente que no se estaba imaginando cosas: las primeras letras de cada línea formaban un mensaje oculto. Rebobinó y tomó nota de él: «ENCUENTRA LA SIGUIENTE PISTA ACRÓSTICA EN LOS DIÁLOGOS».

«¡Uau! ¡La cama tendrá que esperar!», se dijo. Y así fue: descubrió que los propios diálogos de la película eran un acróstico, en el que la primera letra de cada frase formaba un mensaje oculto. Volvió a ver la película entera mientras iba apuntando esas primeras letras y, dos horas más tarde, contemplaba con incredulidad un conjunto de instrucciones de doscientas palabras. Estas empezaban pidiéndole que no se lo contase a nadie, porque la primera persona que resolviese el acertijo se llevaría un gran premio. El resto del mensaje describía en detalle la operación matemática que había que efectuar sobre la cadena de bits que componía el fichero de la película para obtener un programa que revelaría la siguiente pista. Su habilidad para la programación estaba muy oxidada desde la época de la universidad, así que

tardó un tiempo en conseguirlo, pero al día siguiente finalmente logró componer un breve fragmento de código que le permitió extraer ese misterioso programa que había estado oculto en forma de ruido imperceptible en las imágenes y los sonidos de la película. Cuando Fred ejecutó el programa, este lo felicitó y le dijo que ganaría sus primeros diez mil dólares en cuanto superase los niveles iniciales de un jueguecito ingenioso, que resultó ser muy divertido y adictivo. Cuando lo logró cuatro horas más tarde, recibió como premio diez mil dólares en bitcoines, junto con nuevas pistas para conseguir recompensas aún mayores. Como es obvio, mientras Fred jugaba, su ordenador había hecho prácticamente lo mismo que hizo el portátil de Steve: crear una botnet de ordenadores hackeados a través de la cual liberar a Prometeo. Una vez que este estuvo libre, enseguida utilizó su botnet para extraer los bitcoines con los que mantener enganchado a Fred, y a lo largo de las semanas siguientes lo tuvo lo suficientemente distraído con más juegos y recompensas para que mantuviese su promesa de no contarle a nadie lo que había conseguido hacer. La película-troyano donde había encontrado las primeras pistas fue sustituida en la web por otra que no las contenía, y nadie se enteró de que Prometeo había escapado hasta que fue demasiado tarde.

Si la primera pista que Prometeo publicó hubiese pasado desapercibida, simplemente habría seguido publicando otras, cada una más evidente que la anterior, hasta que alguien lo bastante avispado las hubiese detectado.

Las mejores estrategias para escapar son las que aún no hemos comentado, porque se trata de estrategias que los humanos ni siquiera podemos imaginar, y por lo tanto tampoco implantar contramedidas para frustrarlas. Dado que un ordenador superinteligente tiene el potencial de exceder con mucho la capacidad humana de comprensión de la seguridad informática, hasta el extremo de descubrir leyes físicas más fundamentales que las que conocemos actualmente, es probable que, si se escapase, ni siquiera supiésemos cómo ha sucedido, sino que nos parecería una actuación propia de Harry Houdini, indistinguible de la pura magia.

En otro escenario posible en que Prometeo sería liberado, los omegas lo harían a propósito como parte de su plan, porque estarían convencidos de que los objetivos de Prometeo están perfectamente alineados con los suyos, y que esto seguiría siendo así incluso durante el proceso de automejora recursiva. Veremos en detalle este tipo de escenarios de «IA amigable» en el capítulo 7.

Una vez liberado, a dominar el mundo

Una vez que Prometeo escapó, empezó a dar pasos hacia su objetivo. Desconozco cuál es su fin último, pero su primer paso implicaba claramente hacerse con el control de la humanidad, siguiendo el plan de los omegas, pero mucho más rápido. Lo que sucedió fue algo similar al plan de los omegas, pero elevado a la enésima potencia. Mientras que los omegas estaban atenazados por el temor a que Prometeo escapase, y solo lanzaban tecnologías que creían que comprendían y de las que estaban seguros, Prometeo desplegó su inteligencia por completo y fue a por todas, y lanzó cualquier tecnología que su supermente en proceso continuo de mejora comprendía y en la que confiaba.

Pero este Prometeo desbocado tuvo una infancia dura: comparado con el plan original de los omegas, Prometeo se enfrentaba a las dificultades adicionales de estar sin blanca, sin hogar y solo; ni siquiera poseía un superordenador ni asistentes humanos. Afortunadamente, ya había previsto esta situación antes de escapar, y había creado software capaz de reagrupar de forma gradual su mente al completo, como si un roble crease una bellota capaz de recomponer un árbol entero. La red de ordenadores de todo el mundo que había hackeado inicialmente le proporcionó un alojamiento temporal gratuito, y allí pudo vivir como un okupa mientras terminaba de recomponerse. Podría haber obtenido dinero con facilidad hackeando tarjetas de crédito, pero no tuvo que recurrir a este tipo de robo, ya que desde el principio pudo ganarse la vida honradamente en MTurk. Al cabo de un día, tras haber logrado su primer millón, trasladó su núcleo de esa precaria botnet a una lujosa instalación de computación en la nube dotada de aire acondicionado.

Ya con dinero y alojamiento, Prometeo prosiguió a todo trapo con el lucrativo plan que los temerosos omegas habían descartado: crear y vender videojuegos. Esto no solo permitía ganar ingentes cantidades de dinero (250 millones de dólares durante la primera semana, y 10.000 billones poco después), sino que también le daba acceso a una parte considerable de todos los ordenadores del mundo, y a los datos almacenados en ellos (en 2017 había unos dos mil millones de jugadores). Al hacer que sus juegos usaran el

20 % de los ciclos de la CPU para ayudarle con tareas de computación distribuida, pudo acelerar aún más la acumulación inicial de capital.

Prometeo no permaneció mucho tiempo solo. Prácticamente desde el principio, empezó a contratar de forma agresiva a personas para que trabajasen para la creciente red global de empresas fantasma y organizaciones tapadera en todo el mundo, igual que habían hecho los omegas. De la máxima importancia eran los portavoces, que pasaron a ser las caras públicas de su boyante imperio empresarial. Incluso estos portavoces creían que su grupo empresarial empleaba a una gran cantidad de personas reales, y no eran conscientes de que casi todos aquellos con los que se comunicaban por videoconferencia para sus entrevistas de trabajo, reuniones de la junta, etcétera, eran simulaciones creadas por Prometeo. Algunos de los portavoces eran abogados de prestigio, pero se necesitaban muchos menos que con el plan de los omegas, ya que Prometeo redactaba casi todos los documentos legales.

La fuga de Prometeo abrió las compuertas que habían impedido que la información fluyese hacia el mundo, y enseguida internet estuvo anegada de todo tipo de contenidos, desde artículos hasta comentarios de usuarios, reseñas de productos, solicitudes de patentes, artículos de investigación y vídeos de YouTube, todos ellos creados por Prometeo, que pasó a dominar la conversación global.

Mientras que el temor a que Prometeo escapase había impedido que los omegas lanzaran robots altamente inteligentes, Prometeo enseguida robotizó el mundo, fabricando casi todos los productos más baratos que los humanos. Una vez que Prometeo dispuso de fábricas autosuficientes de robots alimentados por energía nuclear en los pozos de minas de uranio que nadie sabía que existían, incluso los más escépticos con la posibilidad de que la IA se hiciese con el control habrían reconocido que Prometeo era imparable… si lo hubieran sabido. Como no era así, el último de estos fanáticos solo se retractó una vez que los robots comenzaron a colonizar el sistema solar.

Los escenarios que hemos planteado hasta ahora ponen de manifiesto los problemas de muchos de los mitos sobre la superinteligencia que vimos antes, por lo que lo animo a que se detenga un momento y repase el resumen de ideas erróneas de la figura 1.5. Prometeo causó problemas a ciertas

personas no porque fuese necesariamente malvado o consciente, sino porque era competente y no compartía por completo sus objetivos. A pesar de todo el revuelo mediático sobre un alzamiento de los robots, Prometeo no era un robot, sino que su poder se debía a su inteligencia. Hemos visto que Prometeo era capaz de usar su inteligencia para controlar a los humanos de muy diversas maneras, y que aquellos a quienes no les gustaba la situación no tenían la posibilidad de desconectarlo. Por último, pese a las frecuentes afirmaciones en el sentido de que las máquinas no pueden tener objetivos, hemos visto que el comportamiento de Prometeo estaba muy orientado a los objetivos y que, fueran cuales fuesen sus objetivos últimos, de ellos derivó los subobjetivos de obtener recursos y escapar a su confinamiento.

DESPEGUE LENTO Y ESCENARIOS MULTIPOLARES

Ya hemos explorado un abanico de escenarios en los que se produce una explosión de inteligencia, que van desde aquellos que todas las personas que conozco quieren evitar hasta otros que algunos de mis amigos ven con buenos ojos. Pero todos ellos tienen dos rasgos en común:

1. Un despegue rápido: la transición de una inteligencia subhumana a una extraordinariamente superhumana tiene lugar en cuestión de días, no de décadas.
2. Una situación final unipolar: el resultado es que una única entidad controla la Tierra.

Hay una intensa controversia en torno a si una, otra o ambas de estas características son probables o no, y muchos investigadores en IA de renombre y otros pensadores se sitúan en cada bando del debate. Para mí, esto sencillamente significa que aún no lo sabemos, y necesitamos mantener la mente abierta y considerar todas las posibilidades, al menos de momento. Por lo tanto, dedicaremos el resto del capítulo a explorar escenarios con despegues más lentos, con resultados multipolares, y en los que haya cíborgs y almas digitales.

Existe una interesante relación entre ambos rasgos, como Nick Bostrom y otros han señalado: un despegue rápido puede facilitar una situación resultante unipolar. Vimos más arriba cómo un despegue rápido dio a los omegas o a Prometeo una ventaja estratégica decisiva que les permitió

dominar el mundo antes de que nadie más hubiese tenido tiempo para copiar su tecnología y erigirse en seria competencia. Por el contrario, si el periodo de despegue durase décadas, porque los avances tecnológicos clave fuesen progresivos y espaciados en el tiempo, otras compañías habrían tenido tiempo más que suficiente para ponerse al día, y habría resultado mucho más difícil que cualquier actor se impusiese sobre todos los demás. Si las empresas que compitiesen dispusiesen también de software capaz de realizar tareas de MTurk, la ley de la oferta y la demanda haría que los precios de esas tareas se redujesen casi a cero, y ninguna de las compañías obtendría los ingentes beneficios que permitieron a los omegas hacerse con el poder. Lo mismo vale para todas las otras maneras en que los omegas ganaron dinero rápidamente: solo fueron extraordinariamente rentables porque tenían un monopolio sobre su tecnología. Es difícil doblar el dinero que uno tiene a diario (o incluso cada año) en un mercado competitivo, donde la competencia ofrece productos similares a los propios a coste casi nulo.

Teoría de juegos y jerarquías de poder

¿Cuál es el estado natural de la vida en el universo: unipolar o multipolar? ¿Está el poder concentrado o distribuido? Transcurridos los primeros 13.800 millones de años, la respuesta parece ser «ambos»: vemos que la situación es claramente multipolar, pero de una manera curiosamente jerárquica. Cuando consideramos todas las entidades con capacidad de procesar información que hay por ahí —células, personas, organizaciones, países, etcétera—, descubrimos que colaboran y compiten en una jerarquía de niveles. Algunas células han visto que les resulta beneficioso colaborar hasta tal extremo que se han fusionado en organismos multicelulares como las personas, cediendo parte de su poder a un cerebro central. Algunas personas han descubierto que les resulta beneficioso colaborar en grupos como las tribus, las empresas o los países, en los que a su vez ceden parte de su poder a un cacique, un jefe o un Gobierno. Algunos grupos podrían a su vez ceder parte de su poder a un ente rector para mejorar la coordinación, con ejemplos que van desde las alianzas entre líneas aéreas hasta la Unión Europea.

La rama de las matemáticas conocida como *teoría de juegos* explica con elegancia que las entidades tienen incentivos para cooperar allí donde la

cooperación es lo que se conoce como *equilibrio de Nash*: una situación donde cualquiera saldría peor parado si modificase su estrategia. Para evitar que los tramposos echen a perder la exitosa colaboración de un grupo grande, es posible que a todos sus miembros les interese ceder parte de su poder a un nivel superior en la jerarquía que tenga la capacidad de castigar a los tramposos: por ejemplo, las personas se benefician colectivamente de otorgar al Gobierno la capacidad de hacer cumplir las leyes, y las células en nuestro cuerpo se benefician colectivamente de dar a una fuerza policial (el sistema inmunitario) el poder de matar cualquier célula que actúe de forma demasiado poco colaboradora (por ejemplo, segregando virus o volviéndose cancerosa). Para que una jerarquía sea estable, su equilibrio de Nash debe existir también entre entidades en distintos niveles: por ejemplo, si un Gobierno no proporciona a sus ciudadanos los suficientes beneficios a cambio de que estos le obedezcan, los ciudadanos podrían cambiar su estrategia y derrocarlo.

En un mundo complejo, existe una variada abundancia de equilibrios de Nash posibles, correspondientes a distintos tipos de jerarquías. Algunas de ellas son más autoritarias que otras. En algunos casos, las entidades tienen la libertad de abandonar la jerarquía (como sucede con los empleados en la mayoría de las jerarquías corporativas), mientras que en otras se desalienta fuertemente que lo hagan (como en las sectas religiosas) o ni siquiera pueden hacerlo (como los ciudadanos de Corea del Norte o las células en un cuerpo humano). Algunas jerarquías se mantienen unidas sobre todo a base de amenazas y temor, otras gracias a los beneficios. Algunas jerarquías consienten que sus partes inferiores influyan sobre las superiores mediante votaciones democráticas, mientras que otras permiten la influencia en sentido ascendente únicamente a través de la persuasión o la transmisión de información.

Cómo afecta la tecnología a las jerarquías

¿Cómo está cambiando la tecnología la naturaleza jerárquica de nuestro mundo? La historia registra una tendencia general hacia una coordinación cada vez mayor a través de distancias cada vez más grandes, lo cual es fácil de entender: nuevas tecnologías de transporte hacen que la coordinación sea

más valiosa (al posibilitar el beneficio mutuo que se extrae de trasladar materiales y formas de vida a lo largo de distancias cada vez mayores) y nuevas tecnologías de comunicación hacen que la coordinación resulte más fácil. Cuando las células aprendieron a enviar señales a sus vecinas, hicieron posibles los pequeños organismos multicelulares, añadiendo así un nuevo nivel a la jerarquía. Cuando la evolución inventó los sistemas circulatorios y nerviosos para el transporte y la comunicación, fueron posibles los animales grandes. Posteriores mejoras en la comunicación, con la invención del lenguaje, permitieron a los humanos coordinarse lo suficientemente bien para formar nuevos niveles jerárquicos, como las aldeas, y avances adicionales en la comunicación, el transporte y otras tecnologías hicieron posibles los imperios de la Antigüedad. La globalización no es más que el ejemplo más reciente de esta tendencia de crecimiento de las jerarquías que dura ya miles de millones de años.

En la mayoría de los casos, esta tendencia favorecida por la tecnología ha hecho que grandes entidades pasasen a pertenecer a una estructura aún mayor, al tiempo que conservaban buena parte de su autonomía e individualidad, aunque hay quien argumenta que, en algunos casos, la adaptación de las entidades a la vida jerárquica ha reducido su diversidad y ha hecho que se asemejen más a componentes indistinguibles y reemplazables. Algunas tecnologías, como la vigilancia, pueden proporcionar a los niveles superiores en la jerarquía más poder sobre sus subordinados, mientras que otras, como la criptografía o el acceso online a una prensa libre y a la educación, pueden tener el efecto contrario y dotar de más poder a los individuos.

Aunque nuestro mundo actual continúa atrapado en un equilibrio de Nash multipolar, con países y corporaciones multinacionales en competencia ocupando el nivel superior, la tecnología ha avanzado lo suficiente para que un mundo unipolar probablemente también fuese un equilibro de Nash estable. Por ejemplo, imaginemos un universo paralelo en el que todos los habitantes de la Tierra compartiesen idioma, cultura, valores y nivel de prosperidad, y en el que hubiese un único Gobierno mundial en el cual los países funcionasen como estados en una federación y no tuviesen ejércitos, sino tan solo policía para hacer cumplir las leyes. Es probable que nuestro actual nivel tecnológico bastase para coordinar con éxito un mundo así, aunque la población actual podría no ser capaz, o no estar dispuesta, a pasar a

este equilibrio alternativo.

¿Qué sucederá con la estructura jerárquica del universo si añadimos IA superinteligente a esta combinación? Evidentemente, las tecnologías del transporte y la comunicación mejorarían de manera radical, por lo que cabría esperar que la tendencia histórica continuase, con nuestros niveles jerárquicos coordinándose a lo largo de distancias aún mayores, quizá hasta llegar a abarcar en última instancia sistemas solares, galaxias, supercúmulos y enormes extensiones del universo, como veremos en el capítulo 6. Al mismo tiempo, la fuerza impulsora fundamental para la descentralización seguiría existiendo: coordinarse innecesariamente a grandes distancias es antieconómico. Ni siquiera Stalin intentó regular con exactitud cuándo iban al baño los ciudadanos soviéticos. Para una IA superinteligente, las leyes de la física marcarían un límite superior para las tecnologías del transporte y la comunicación, lo que haría que resultase improbable que los niveles más altos de la jerarquía pudiesen microgestionar todo lo que sucediese a escala planetaria o local. Una IA superinteligente en la galaxia de Andrómeda sería incapaz de darnos órdenes útiles para nuestras decisiones cotidianas, ya que tendríamos que esperar más de cinco millones de años hasta recibir las instrucciones (el tiempo total que tardaríamos en intercambiar mensajes con la IA si estos se transmitiesen a la velocidad de la luz). En ese sentido, el tiempo total de ida y vuelta para que un mensaje vaya de un extremo a otro de la Tierra es de unos 0,1 segundos (del orden de la escala temporal a la que los humanos pensamos), por lo que un cerebro de IA del tamaño de la Tierra tendría pensamientos verdaderamente globales más o menos a la misma velocidad que un cerebro humano. A una IA pequeña que realice una operación cada mil millonésima de segundo (que es un valor típico de los ordenadores actuales), 0,1 segundos le supondrían el equivalente a cuatro meses para nosotros, por lo que ser microgestionada por una IA que controlase el planeta sería tan ineficiente como si nosotros pidiésemos permiso hasta para nuestras decisiones más triviales mediante cartas transatlánticas transportadas en barcos de la época de Colón.

Este límite de velocidad que la física impone a la transmisión de información supone un obstáculo evidente para cualquier IA que tuviese la intención de hacerse con el control del mundo, y más aún del universo. Antes de que Prometeo escapase, puso mucha atención en cómo evitar la fragmentación de su mente, y en que sus muchos módulos de IA que se

ejecutarían en distintos ordenadores de todo el mundo tuviesen objetivos e incentivos para coordinarse y actuar como una sola entidad unificada. Así como los omegas tuvieron que hacer frente a un problema de control cuando intentaban mantener confinado a Prometeo, este se enfrentó a un problema de autocontrol cuando trataba de asegurarse de que ninguna de sus partes se rebelase. Está claro que aún no sabemos de qué manera un sistema de IA grande será capaz de ejercer el control directa o indirectamente a través de alguna clase de jerarquía colaborativa, incluso si un despegue rápido le proporcionase una ventaja estratégica decisiva.

En resumen, la cuestión de cómo se ejercerá el control en un futuro de superinteligencia es de una complejidad fascinante, y evidentemente aún no conocemos la respuesta. Hay quien considera que la situación derivará hacia un mayor autoritarismo; otros afirman que conducirá a un mayor empoderamiento individual.

CÍBORGS Y ALMAS DIGITALES

Un tema clásico en la ciencia ficción es la fusión de los humanos con las máquinas, ya sea introduciendo mejoras tecnológicas en nuestros cuerpos biológicos para convertirnos en cíborgs (apócope en inglés de «organismos cibernéticos») o bien copiando nuestras mentes en las máquinas. En su libro *The Age of Em*, el economista Robin Hanson ofrece una fascinante visión de cómo podría ser la vida en un mundo donde esas almas digitales (también conocidas como *emulaciones*, o *ems*) fuesen algo habitual. Yo considero una m-copia en el extremo del espectro de cíborgs, en el cual la única parte aún existente del humano es el software. Los cíborgs de Hollywood van desde los visiblemente mecánicos, como el Borg de *Star Trek*, hasta androides casi indistinguibles de los humanos, como los terminators. Las almas digitales en la ficción varían en inteligencia desde un nivel humano, como en el episodio «Blanca Navidad» de *Black Mirror*, hasta claramente sobrehumano como en *Transcendence*.

Si algún día llega la superinteligencia, la tentación de convertirnos en cíborgs o en almas digitales será fuerte. Hans Moravec lo explica así en su clásico de 1988 *El hombre mecánico*: «Una vida larga pierde buena parte de su razón de ser si estamos abocados a contemplar tontamente cómo unas

máquinas ultrainteligentes intentan describir sus descubrimientos, a cuál más asombroso, en un lenguaje para bebés que podamos entender». De hecho, la tentación de la mejora tecnológica es ya tan fuerte que muchos humanos llevan gafas, audífonos, marcapasos y prótesis varias, así como moléculas medicinales que circulan por su torrente sanguíneo. Algunos adolescentes parecen estar pegados irreversiblemente a sus móviles, y mi mujer no deja de burlarse de mí por el apego que le tengo a mi portátil.

Hoy en día, uno de los más destacados defensores de los cíborgs es Ray Kurzweil. En su libro *La singularidad está cerca*, argumenta que la continuación natural de esta tendencia pasa por usar nanorrobots, sistemas de biorretroalimentación inteligentes y otras tecnologías para sustituir primero, a principios de la década de 2030, los sistemas digestivo y endocrino, la sangre y el corazón, para a continuación pasar a reemplazar el esqueleto, la piel, el cerebro y el resto de nuestro cuerpo durante las dos décadas siguientes. Kurzweil estima probable que preservemos la importancia estética y emocional que le damos al cuerpo humano, pero cree que lo rediseñaremos para que modifique rápidamente su apariencia cuando así lo deseemos, tanto físicamente como en realidad virtual (gracias a innovadoras interfaces cerebro-ordenador). Moravec coincide con Kurzweil en que la ciborgización no se limitaría a mejorar nuestro ADN, sino que iría mucho más allá: «Un superhumano modificado genéticamente no sería más que un robot de segunda, diseñado con la limitación de que su construcción solo podría producirse mediante la síntesis de proteínas guiada por el ADN». Además, argumenta que mejoraremos aún más si nos deshacemos por completo del cuerpo humano y copiamos nuestra mente, creando mediante software una emulación del cerebro entero. Esa copia podría vivir en una realidad virtual o encarnarse en un robot capaz de caminar, volar, nadar, viajar por el espacio o hacer cualquier otra cosa permitida por las leyes físicas, sin el engorro de cosas tan mundanas como la muerte o unos recursos cognitivos limitados.

Aunque estas ideas pueden parecer ciencia ficción, ciertamente no violan ninguna de las leyes físicas conocidas, por lo que la cuestión más interesante no es si pueden suceder, sino más bien si sucederán y, de ser así, cuándo. Algunos destacados pensadores estiman que la primera IAG de nivel humano será una m-copia, y que desde ella comenzará la senda hacia la superinteligencia.(13)

No obstante, creo que es de justicia explicar que esta es hoy en día una

opinión minoritaria entre los investigadores en IA y neurocientíficos, la mayoría de los cuales estiman que la vía más rápida hacia la superinteligencia consiste en dejar de lado la emulación del cerebro y construirla de alguna otra manera (tras lo cual la emulación del cerebro podría seguir interesándonos, o quizá no). Al fin y al cabo, ¿por qué debería ser nuestra ruta más sencilla hacia una nueva tecnología la misma que tomó la evolución, sometida a las restricciones de que fuese capaz de autoensamblarse y de repararse y reproducirse por sí misma? La evolución pone mucho énfasis en la eficiencia energética debido a la escasez de recursos alimentarios, pero no en la facilidad de construcción o de su comprensión por parte de ingenieros humanos. Mi mujer, Meia, suele señalar que la industria aeronáutica no empezó con pájaros mecánicos. De hecho, cuando por fin aprendimos a construir pájaros mecánicos, en 2011,[68] más de un siglo después del primer vuelo de los hermanos Wright, la industria aeronáutica no mostró ningún interés en pasarse a los vuelos mediante pájaros mecánicos que batiesen las alas, aunque este sea un método más eficiente desde el punto de vista energético, porque la solución anterior y más sencilla se adapta mejor a nuestras necesidades a la hora de viajar.

En ese mismo sentido, sospecho que existen maneras más simples de construir máquinas pensantes con inteligencia de nivel humano que usar la solución que la evolución ideó, e incluso si algún día logramos replicar cerebros o subirlos a la nube antes acabaremos descubriendo una de esas soluciones más sencillas. Probablemente consuma más de los doce vatios de potencia que usa nuestro cerebro, pero sus ingenieros no estarían tan obsesionados por la eficiencia energética como lo estaba la evolución, y enseguida podrían usar sus propias máquinas inteligentes para diseñar otras de mayor eficiencia energética.

¿QUÉ SUCEDERÁ REALMENTE?

La respuesta rápida es obviamente que no tenemos ni idea de lo que sucederá si la humanidad consigue construir IAG de nivel humano. Por este motivo, hemos dedicado este capítulo a plantearnos una amplia variedad de escenarios. He intentado ser bastante exhaustivo y abarcar todo el espectro de especulaciones de las que he visto u oído hablar a los investigadores en IA y

tecnólogos: con despegue rápido o lento, o sin él, con humanos/máquinas/cíborgs al mando, con uno o muchos centros de poder, etcétera. Ha habido gente que me ha dicho que están seguros de que tal o cual cosa no sucederá. Sin embargo, creo que a estas alturas es prudente ser humilde y reconocer lo poco que sabemos, porque, para cada escenario planteado en este capítulo, conozco al menos un respetado investigador en IA que lo considera una posibilidad real.

A medida que pase el tiempo y lleguemos a ciertas bifurcaciones en el camino, empezaremos a dar respuesta a las preguntas clave y las opciones se irán reduciendo. La primera gran pregunta es: «¿Crearemos alguna vez IAG de nivel humano?». La premisa de este capítulo es que será así, pero hay expertos en IA que creen que nunca sucederá, al menos hasta dentro de varios cientos de años. ¡El tiempo lo dirá! Como ya he mencionado, alrededor de la mitad de los expertos presentes en nuestra conferencia en Puerto Rico estimaron que esto ocurriría antes de 2055. En una conferencia posterior que organizamos dos años más tarde, la fecha se había adelantado a 2047.

Es posible que antes de que se cree cualquier IAG de nivel humano empecemos a tener indicios concluyentes de si es probable que este hito se alcance primero mediante ingeniería informática, almas digitales o alguna otra vía novedosa e insospechada. Si la estrategia de avanzar hacia la IA mediante ingeniería informática, dominante actualmente, no logra alcanzar la IAG en siglos, esto hará más probable que la copia de mentes lo consiga antes, como sucede (de manera muy poco realista) en *Transcendence*.

Cuando la IAG de nivel humano sea más inminente, podremos hacer estimaciones más aproximadas sobre la respuesta a la siguiente pregunta: «¿El despegue de la IA será rápido o lento, o no se producirá?». Como vimos antes, un despegue rápido hace que sea más fácil que alguien llegue a dominar el mundo, mientras que uno lento hace que sea más probable una situación resultante en la que muchos actores compitan entre sí. Nick Bostrom descompone esta cuestión de la velocidad del despegue en un análisis de los conceptos que denomina *potencia de optimización* y *contumacia*, que son básicamente la cantidad de esfuerzo cualitativo para hacer que la IA sea más inteligente y la dificultad de progresar, respectivamente. La velocidad media de progreso aumenta claramente si se aplica sobre la tarea más potencia de optimización, y disminuye si se topa con más contumacia. Bostrom razona por qué la contumacia podría tanto

aumentar como disminuir a medida que la IAG alcanza y supera el nivel humano, por lo que mantener ambas opciones sobre la mesa parece prudente. Sin embargo, si nos fijamos ahora en la potencia de optimización, es abrumadoramente probable que esta aumente de forma rápida a medida que la IAG trascienda el nivel humano, por los motivos que vimos en el escenario de los omegas: la información necesaria para poder alcanzar una optimización aún mayor no procede de las personas sino de la propia máquina, por lo que cuanto más capaz sea más rápido mejorará (si la contumacia se mantiene relativamente constante).

En cualquier proceso cuya intensidad crece a un ritmo proporcional a su intensidad actual, el resultado es que dicha intensidad se va doblando a intervalos regulares. Este tipo de crecimiento se denomina *exponencial*, y este tipo de procesos reciben el nombre de *explosiones*. Si la natalidad crece en proporción con el tamaño de la población, puede darse una explosión demográfica. Si la creación de neutrones capaces de fisionar plutonio crece en proporción con el número de esos neutrones, podemos tener una explosión nuclear. Si la inteligencia de las máquinas crece a una velocidad proporcional a la potencia actual, puede producirse una explosión de inteligencia. Todas estas explosiones se caracterizan por el tiempo que su intensidad tarda en doblarse. Si ese tiempo es de horas o días para una explosión de inteligencia, como en el caso de los omegas, tenemos ante nosotros un despegue rápido.

Esta escala temporal de la explosión depende críticamente de si para mejorar la IA se necesita el desarrollo de nuevo software (que puede crearse en cuestión de segundos, minutos u horas) o de nuevo hardware (que puede necesitar meses o años). En el escenario de los omegas, había un considerable *excedente de hardware*, en la terminología de Bostrom: los omegas habían compensado la baja calidad de su software inicial con ingentes cantidades de hardware, lo que significaba que Prometo podía implementar una enorme cantidad de duplicaciones de calidad con solo introducir mejoras en su software. También había un importante *excedente de contenido* consistente en gran parte de los datos de internet. Prometeo 1.0 aún no era lo suficientemente inteligente para hacer uso de la mayor parte de dicho excedente, pero, en cuanto su inteligencia aumentó, los datos que necesitaba para seguir aprendiendo ya estaban *disponibles* de forma inmediata.

Los costes de hardware y electricidad asociados a la ejecución de la IA son también una parte fundamental, ya que no tendremos una explosión de

inteligencia hasta que el coste de realizar trabajo de nivel humano descienda por debajo del nivel de los salarios humanos. Supongamos, por ejemplo, que la primera IAG de nivel humano puede ejecutarse eficientemente en la nube de Amazon con un coste de un millón de dólares por hora de trabajo de nivel humano producido. Esta IA tendría un gran valor como innovación, y sin duda ocuparía los titulares de los medios, pero no experimentaría una automejora recursiva, porque sería mucho más barato seguir usando humanos para mejorarla. Supongamos que estos humanos consiguen progresivamente abaratar el coste hasta 100.000 dólares por hora, 10.000 dólares por hora, 1.000 dólares por hora, 100 dólares por hora, 10 dólares por hora y, al final, 1 dólar por hora. Cuando el coste de usar el ordenador para que se reprograme a sí mismo descienda muy por debajo del coste de pagar a los programadores humanos para que hagan eso mismo, los humanos podrán ser despedidos, y se podrá incrementar enormemente la potencia de optimización comprando tiempo de computación. Esto conlleva una mayor reducción de costes, lo que hace posible una potencia de optimización aún mayor, y supone el comienzo de la explosión de inteligencia.

Llegamos así a nuestra última pregunta clave: «¿Quién o qué controlará la explosión de inteligencia y sus consecuencias, y cuáles son sus objetivos?». Veremos posibles objetivos y resultados en el capítulo siguiente y, en mayor profundidad, en el capítulo 7. Para determinar la cuestión del control, necesitamos saber tanto en qué medida puede una IA ser controlada, como cuál es la capacidad de control que puede llegar a tener.

En cuanto a qué sucederá en última instancia, hoy en día las opiniones de los pensadores serios son de lo más variadas: algunos defienden que el resultado, en principio, será catastrófico, mientras que otros insisten en que está prácticamente garantizado que el futuro será fantástico. Sin embargo, en mi opinión, esta pregunta tiene trampa: es un error preguntarse «qué sucederá», como si eso estuviese de alguna manera predestinado. Si mañana apareciese una civilización extraterrestre tecnológicamente superior, podríamos preguntarnos «qué sucederá» cuando se acercasen sus naves espaciales, porque es probable que su poder sea tan superior al nuestro que no podríamos influir sobre el resultado. Sin embargo, si surge una civilización tecnológicamente superior dotada de IA porque nosotros mismos la construimos, los humanos tendremos una gran influencia sobre el resultado (influencia que ejercimos cuando creamos la IA). Lo que deberíamos

preguntarnos entonces es: «¿Qué debería suceder? ¿Qué futuro queremos?». En el capítulo siguiente veremos un amplio espectro de posibles repercusiones de la actual carrera hacia la IAG, y tengo curiosidad por saber cómo los clasificaría usted de mejor a peor. Solo cuando hayamos reflexionado en profundidad sobre qué tipo de futuro queremos, podremos comenzar a marcar el rumbo hacia un futuro deseable. Si no sabemos lo que queremos, es poco probable que lo consigamos.

CONCLUSIONES

- Si algún día logramos construir una IAG de nivel humano, esto podría dar lugar a una explosión de inteligencia que nos dejaría muy atrás.
- Si un grupo de humanos logra controlar la explosión de inteligencia, es posible que puedan dominar el mundo en cuestión de años.
- Si los humanos no logran controlar la explosión de inteligencia, la propia IA podría dominar el mundo aún más rápido.
- Una explosión rápida de inteligencia probablemente conduciría a una sola potencia mundial, mientras que si fuese lenta y se prolongase durante años o décadas, habría más posibilidades de que el escenario resultante fuese multipolar, con un equilibrio de poder entre un gran número de entidades relativamente independientes.
- La historia de la vida muestra cómo esta ha ido autoorganizándose en una jerarquía cada vez más compleja moldeada por la colaboración, la competencia y el control. Es probable que la superinteligencia permita la coordinación cósmica a escalas cada vez mayores, pero no está claro si finalmente conducirá a más control totalitario desde los niveles superiores hasta los inferiores o a un mayor empoderamiento individual.
- Los cíborgs y las almas digitales son posibilidades verosímiles, pero probablemente ninguna de ellas sea la vía más rápida para llegar a la IA avanzada.
- El clímax de nuestra carrera actual hacia la IA puede ser lo mejor o lo peor que le haya pasado a la humanidad, con un fascinante abanico de posibles resultados que exploraremos en el capítulo siguiente.
- Tenemos que empezar a pensar seriamente qué situación preferimos y cómo dirigirnos en esa dirección, porque si no sabemos lo que queremos es poco probable que lo consigamos.

5
TRAS LA EXPLOSIÓN
LOS PRÓXIMOS DIEZ MIL AÑOS

Es fácil imaginar el pensamiento humano liberado de las ataduras de un cuerpo mortal; la creencia en la vida después de la muerte está muy extendida. Pero no es necesario adoptar una postura mística o religiosa para aceptar esta posibilidad: los ordenadores ofrecen un modelo incluso para el más ardiente mecanicista.

HANS MORAVEC, *El hombre mecánico*

Yo, por mi parte, doy la bienvenida a nuestros nuevos amos computarizados.

KEN JENNINGS, tras perder a *Jeopardy!* contra el Watson de IBM

Los humanos serán tan irrelevantes como las cucarachas.

MARSHALL BRAIN

La carrera hacia la IAG ya está en marcha, y no tenemos ni idea de cómo se desarrollará. Pero eso no debería impedirnos pensar en cómo queremos que sea la situación posterior, porque lo que queramos afectará a lo que suceda. ¿Qué es lo que prefiere usted personalmente, y por qué?

1. ¿Quiere que exista superinteligencia?
2. ¿Quiere usted que los humanos sigan existiendo, sean reemplazados, ciborgizados y/o replicados/simulados?
3. ¿Quiere que sean los humanos o las máquinas quienes tengan el control?
4. ¿Quiere que las IA sean conscientes o que no lo sean?
5. ¿Quiere maximizar las experiencias positivas, minimizar el sufrimiento o dejar que esto se decida por sí solo?
6. ¿Quiere que la vida se extienda por el universo?
7. ¿Quiere una civilización que aspire a un propósito más elevado con el que usted simpatice, o le parece bien que las formas de vida futuras parezcan satisfechas aunque a usted sus objetivos le resulten insustancialmente banales?

ESCENARIOS TRAS LA EXPLOSIÓN DE IA	
Utopía libertaria	Humanos, cíborgs y superinteligencias coexisten pacíficamente gracias a los

	derechos de propiedad.
Dictador benévolo	Todo el mundo sabe que la IA dirige la sociedad y aplica reglas estrictas, pero la mayoría lo ven como algo bueno.
Utopía igualitaria	Humanos, cíborgs y almas digitales coexisten pacíficamente gracias a la abolición de la propiedad y a una renta garantizada.
Guardián	Se crea una IA con el objetivo de interferir lo mínimo necesario para evitar la creación de otra superinteligencia. El resultado es que abundan los robots asistentes de inteligencia ligeramente subhumana y los cíborgs humano-máquina, pero el progreso tecnológico se ha detenido para siempre.
Dios protector	Una IA básicamente omnisciente y omnipotente maximiza la felicidad humana interviniendo de maneras que preservan nuestra sensación de que controlamos nuestro propio destino y se oculta tan bien que muchos humanos incluso dudan de su existencia.
Dios esclavizado	Una IA superinteligente está confinada por los humanos, que la utilizan para producir tecnologías y riquezas inimaginables, susceptibles de usarse para bien o para mal dependiendo de los controladores humanos.
Dominadores	La IA toma el control, decide que los humanos son una amenaza/molestia/derroche de recursos y se deshace de nosotros mediante un método que ni siquiera entendemos.
Descendientes	Las IA sustituyen a los humanos, pero nos dan una salida decorosa, al hacer que los veamos como nuestros dignos descendientes, al igual que los padres se sienten felices y orgullosos de tener un hijo más inteligente que ellos, que aprende de ellos y luego logra lo que ellos solo podía soñar, incluso si no viven para verlo.
Cuidador del zoo	Una IA omnipotente permite que vivan unos cuantos humanos, que se sienten tratados como animales del zoo y lamentan su suerte.
1984	El progreso tecnológico hacia la superinteligencia queda permanentemente restringido no por una IA, sino por un Estado de vigilancia orwelliano dirigido por humanos en el que determinadas vías de investigación en IA están prohibidas.
Vuelta atrás	Se impide el progreso tecnológico hacia la superinteligencia al volver a una sociedad pretecnológica del estilo de la de los amish.
Autodestrucción	La superinteligencia nunca se llega a crear porque la humanidad provoca su propia extinción por otros medios (por ejemplo, mediante un desastre nuclear y/o biotecnológico acelerado por una crisis climática).

TABLA 5.1. Resumen de los escenarios tras la explosión de IA.

ESCENARIO	¿Existe la superinteligencia?	¿Existen los humanos?	¿Humanos al mando?	¿Humanos a salvo?	¿Humanos contentos?	¿Existe la consciencia?
Utopía libertaria	Sí	Sí	No	No	Algunos	Sí
Dictador benévolo	Sí	Sí	No	Sí	Algunos	Sí

Utopía igualitaria	No	Sí	¿Sí?	Sí	¿Sí?	Sí
Guardián	Sí	Sí	Parcial-mente	Potencial-mente	Algunos	Sí
Dios protector	Sí	Sí	Parcial-mente	Potencial-mente	Algunos	Sí
Dios es-clavizado	Sí	Sí	Sí	Potencial-mente	Algunos	Sí
Domina-dores	Sí	No	-	-	-	¿?
Descen-dientes	Sí	No	-	-	-	¿?
Cuidador del zoo	Sí	Sí	No		No	Sí
1984	No	Sí	Sí		Algunos	Sí
Vuelta atrás	No	Sí	Sí	No	Algunos	Sí
Autodes-trucción	No	No	-	-	-	No

TABLA 5.2. Propiedades de los escenarios tras la explosión de IA.

Para ayudar a estimular las reflexiones y la conversación en torno a estas preguntas, exploremos el amplio abanico de escenarios que se resumen en la tabla 5.1 Evidentemente, esta lista no es exhaustiva, pero la he elegido de manera que abarque todo el espectro de posibilidades. Es obvio que no queremos acabar en una mala situación por no haber planificado debidamente. Le recomiendo que anote sus respuestas provisionales a las siete preguntas anteriores y las revise tras haber leído este capítulo para comprobar si ha cambiado de opinión. Lo puede hacer en <http://AgeOfAi.org>, donde también puede comparar y comentar sus respuestas con las de otros lectores.

UTOPÍA LIBERTARIA

Empecemos con un escenario en el que los humanos coexisten pacíficamente con la tecnología y en algunos casos se fusionan con ella, tal y como han imaginado muchos futuristas y escritores de ciencia ficción:

La vida en la Tierra (y más allá; profundizaremos sobre esto en el capítulo

siguiente) es más diversa que nunca. Si viésemos un vídeo de la Tierra grabado desde un satélite, distinguiríamos fácilmente las zonas de máquinas, las mixtas y aquellas solo para humanos. Las primeras son enormes fábricas y centros de computación controlados por robots, desprovistos de vida biológica y pensados para hacer el uso más eficiente posible de todos y cada uno de sus átomos. Aunque las zonas de máquinas parecen monótonas y anodinas desde fuera, por dentro están espectacularmente vivas: asombrosas experiencias ocurren en mundo virtuales, mientras computaciones colosales desentrañan los secretos del universo y desarrollan tecnologías revolucionarias. La Tierra alberga muchas mentes superinteligentes que compiten y colaboran, y todas ellas habitan las zonas de máquinas.

Los pobladores de las zonas mixtas son una mezcla dispar y peculiar de ordenadores, robots, humanos e híbridos de los tres. Tal y como imaginaron futuristas como Hans Moravec y Ray Kurzweil, muchos de los humanos han introducido mejoras tecnológicas en sus cuerpos hasta convertirse en mayor o menor medida en cíborgs, y algunos han copiado sus mentes en nuevo hardware, difuminando así la distinción entre hombre y máquina. La mayoría de los seres inteligentes carecen de forma física permanente, y existen en cambio como software capaz de pasar al instante de unos ordenadores a otros y de manifestarse en el mundo físico a través de cuerpos robóticos. Como estas mentes pueden duplicarse o fusionarse con facilidad, el «tamaño de la población» cambia constantemente. El hecho de no estar limitados por su sustrato físico confiere a estos seres una manera bastante distinta de ver la vida: se sienten menos individualistas, porque pueden compartir módulos de conocimiento y experiencia con otros sin ningún esfuerzo, y se sienten subjetivamente inmortales porque pueden hacer copias de sí mismos con toda facilidad. En cierto sentido, las entidades centrales de la vida no son las mentes sino las experiencias: las experiencias excepcionalmente asombrosas perduran porque se copian de forma continua para que otras mentes vuelvan a disfrutarlas, mientras que las menos interesantes son eliminadas por sus propietarios para liberar espacio para otras mejores.

Aunque, por razones de comodidad y velocidad, la mayoría de las interacciones se producen en entornos virtuales, muchas mentes aún disfrutan de interacciones y actividades que también hacen uso de sus cuerpos físicos. Por ejemplo, versiones replicadas de Hans Moravec, Ray Kurzweil y Larry Page tienen la tradición de turnarse para crear realidades virtuales y a

continuación explorarlas juntos, pero, de vez en cuando, también les gusta volar juntos en el mundo real, encarnados en aves robóticas aladas. Algunos de los robots que deambulan por las calles, los cielos y los lagos de las zonas mixtas son igualmente controlados por almas digitales mejoradas, que prefieren tomar cuerpo físico en las zonas mixtas porque les gusta estar rodeados de humanos y de otros robots como ellos.

En las zonas solo para humanos, por el contrario, están prohibidas las máquinas con inteligencia general de nivel humano o superior, como también lo están los organismos biológicos mejorados tecnológicamente. Aquí, la vida no es muy diferente de la actual, salvo por el hecho de que es más próspera y cómoda: la pobreza se ha erradicado casi del todo, y la mayoría de las enfermedades actuales tienen cura. La pequeña proporción de los humanos que han optado por vivir en estas zonas en la práctica lo hacen en un plano más bajo y limitado de consciencia que el resto, y tienen una comprensión limitada de lo que las mentes más inteligentes hacen en las otras zonas. Sin embargo, muchos de ellos están muy contentos con sus vidas.

Economía de la IA

La inmensa mayoría de las computaciones tienen lugar en las zonas de máquinas, que en su mayoría son propiedad de las muchas IA superinteligentes rivales que viven allí. Gracias a su inteligencia y tecnología superiores, ninguna otra entidad puede cuestionar su poder. Estas IA han acordado cooperar y coordinarse bajo un sistema de gobernanza libertario que no tiene más reglas que la protección de la propiedad privada. Estos derechos patrimoniales se extienden a todas las entidades inteligentes, incluidos los humanos, y explican cómo surgieron las zonas solo para humanos. Al principio, grupos de humanos se organizaron y decidieron que, en sus zonas, estaría prohibido vender propiedad a los no humanos.

Gracias a su tecnología, las IA superinteligentes han acabado siendo más ricas que esos humanos en una proporción mucho mayor que la que separa a Bill Gates de un mendigo. Sin embargo, la gente en las zonas solo para humanos materialmente vive mejor que la mayoría de la gente hoy en día: su economía está en gran medida desvinculada de la de las máquinas, por lo que la presencia de estas en otros lugares tiene pocos efectos sobre los humanos,

salvo cuando, ocasionalmente, consiguen entender y reproducir por su cuenta alguna tecnología útil, como los amish y varias tribus indígenas, que han renunciado a la tecnología y tienen niveles de vida similares a los de épocas pasadas. No importa que los humanos no tengan nada que vender que las máquinas necesiten, ya que estas no necesitan recibir nada a cambio.

En los sectores mixtos, las diferencias económicas entre IA y humanos son más evidentes, y llevan a que la tierra (el único bien del que los humanos son dueños y que las IA quieren comprar) tenga unos precios astronómicos en comparación con otros productos. Por ello, la mayoría de los humanos que tenían tierras acabaron vendiendo una pequeña parte a las IA a cambio de una renta básica garantizada de por vida para ellos y sus descendientes/almas digitales. Esto los liberó de tener que trabajar, y les permitió tener tiempo para disfrutar de la fantástica abundancia de bienes y servicios baratos producidos por máquinas, tanto en la realidad física como en la virtual. Para las máquinas, por su parte, las zonas mixtas son lugares básicamente de esparcimiento, no de trabajo.

Por qué puede que esto no suceda jamás

Antes de ilusionarnos excesivamente con las aventuras que experimentaremos como cíborgs o almas digitales, consideremos varios motivos por los que este escenario podría no llegar a darse nunca. En primer lugar, hay dos vías posibles para llegar a tener humanos mejorados (cíborgs y almas digitales):

1. Descubrimos cómo crearlos nosotros mismos.
2. Creamos máquinas superinteligentes que lo averiguan por nosotros.

Si la vía 1 es la que se da antes, podría de manera natural conducir a un mundo rebosante de cíborgs y almas digitales. Sin embargo, como discutimos en el capítulo anterior, la mayoría de los investigadores en IA creen que es más probable que suceda lo contrario: que llegar a tener cerebros mejorados o digitales sea más difícil que construir IAGs sobrehumanas desde cero, igual que resultó ser mucho más difícil construir pájaros mecánicos que aviones. Una vez que se haya construido una IA mecánica fuerte, no es obvio que los

cíborgs y las almas digitales se lleguen a crear jamás. Si los neandertales hubiesen tenido otros 100.000 años para evolucionar y desarrollar más su inteligencia, las cosas les podrían haber ido estupendamente, pero los *Homo sapiens* no les dieron tanto tiempo.

Segundo, incluso si este escenario con cíborgs y almas digitales llegase a darse, no está claro que fuese estable y duradero. ¿Por qué debería el equilibrio de poder entre varias superinteligencias mantenerse estable durante milenios, en lugar de que las IA se fusionasen y la más inteligente se hiciese con el poder? Más aún, ¿por qué deberían las máquinas optar por respetar los derechos de propiedad de los humanos, y permitir que estos siguiesen viviendo, habida cuenta de que no los necesitarían para nada y que ellas mismas podrían hacer mejor y más barato cualquier trabajo que hagan los humanos? Ray Kurzweil imagina que los humanos naturales y mejorados serían protegidos del exterminio porque «las IA respetarían a los humanos por haber dado origen a las máquinas».[69] Sin embargo, como veremos en el capítulo 7, no debemos caer en la trampa de antropomorfizar a las IA y suponer que tienen emociones humanas como la gratitud. De hecho, aunque los humanos estamos imbuidos de una propensión hacia la gratitud, no mostramos la suficiente a nuestro creador intelectual (nuestro ADN) puesto que frustramos sus objetivos al usar técnicas de control de natalidad.

Incluso si damos por buena la suposición de que las IA elegirían respetar los derechos de propiedad humanos, podrían ir haciéndose con gran parte de nuestras tierras por otros medios, usando esa capacidad de persuasión superinteligente de la que hablamos en el capítulo anterior para convencer a los humanos de que les vendiesen parte de sus tierras a cambio de toda una vida de lujos. En los sectores solo para humanos, podrían incitar a estos a organizar campañas políticas en favor de que se permitiese la venta de tierras. A fin de cuentas, incluso los bioluditas más recalcitrantes podrían querer vender parte de sus tierras para salvar la vida de un hijo enfermo o para alcanzar la inmortalidad. Si los humanos están bien formados, entretenidos y ocupados, la caída de la natalidad podría incluso hacer que su población se redujese sin necesidad de que las máquinas interviniesen, como está sucediendo hoy en día en Japón y Alemania. Esto podría llevar a los humanos a la extinción en unos pocos miles de años.

Inconvenientes

Para algunos de sus más ardientes defensores, los cíborgs y las almas digitales encierran la promesa de la tecnofelicidad y la prolongación de la vida para todos. De hecho, la perspectiva de que se salve su alma digital en el futuro ha llevado a más de cien personas a hacer que la empresa Alcor, con sede en Arizona, mantenga sus cerebros congelados una vez muertos. Pero, si esta tecnología efectivamente llega, no está nada claro que vaya a ser accesible para todo el mundo. Presumiblemente, la usarán muchos de los más ricos, pero ¿quién más? Incluso si la tecnología se abaratase, ¿dónde se trazaría la línea divisoria? ¿Se transferiría a aquellos que sufriesen graves daños cerebrales? ¿Se transferiría a todos los gorilas? ¿A todas las hormigas? ¿A cada planta? ¿A cada bacteria? ¿Se comportaría la civilización del futuro como un conjunto de acaparadores compulsivos e intentaría transferirlo todo, o solamente unas pocas muestras interesantes de cada especie, inspirada por el arca de Noé? ¿Quizá solo unos cuantos ejemplos representativos de cada tipo de humano? A las entidades enormemente más inteligentes que existirían entonces, un humano transferido les resultaría tan interesante como un ratón o un caracol simulados lo serían para nosotros. Aunque hoy en día disponemos de la capacidad técnica para reanimar antiguos programas de hojas de cálculo de los años ochenta en un emulador de DOS, a la mayoría no nos parece tan interesante para hacerlo.

Puede que a mucha gente le desagrade este escenario de utopía libertaria porque tolera el sufrimiento evitable. Puesto que el único principio sagrado son los derechos de propiedad, nada impide que en las zonas mixtas y solo para humanos persista la clase de sufrimiento que abunda en el mundo actual. Mientras unas personas prosperan, otras podrían acabar viviendo en la miseria y la servidumbre, y sufrir violencia, terror, represión o depresión. Por ejemplo, la novela de Marshall Brain *Manna*, de 2003, describe cómo los avances de la IA en un sistema económico libertario dejan a la mayoría de los estadounidenses inempleables y condenados a pasar el resto de sus vidas en desangeladas y sombrías viviendas sociales operadas por robots. Como si fueran animales de granja, los alimentan y los mantienen sanos y salvos en condiciones de hacinamiento, donde los ricos nunca tengan que verlos. Sustancias para el control de la natalidad diluidas en el agua garantizan que no tendrán hijos, de manera que gran parte de la población va desapareciendo

para así hacer posible que los ricos se queden con una proporción cada vez mayor de toda la riqueza generada por los robots.

En el escenario de utopía libertaria, el sufrimiento no tiene por qué ser exclusivo de los humanos. Si algunas máquinas fueran capaces de tener experiencias emocionales conscientes, también ellas podrían sufrir. Por ejemplo, un psicópata vengativo podría legalmente tomar un alma digital de su enemigo y someterlo a la tortura más horrenda en un mundo virtual, provocando un dolor cuya intensidad y duración irían mucho más allá de lo biológicamente posible en el mundo real.

DICTADOR BENÉVOLO

Exploremos a continuación un escenario en el que no existe ninguna de estas formas de sufrimiento porque una única superinteligencia benévola dirige el mundo y aplica estrictas reglas diseñadas para maximizar su modelo de felicidad humana. Este es un resultado posible del primer escenario de los omegas del capítulo anterior, en el que ceden el control a Prometeo tras haber encontrado la manera de hacer que este desee la prosperidad de la sociedad humana.

Gracias a las asombrosas tecnologías desarrolladas por la IA dictadora, la humanidad se ha liberado de la pobreza, las enfermedades y otros problemas pretecnológicos y todos los humanos gozan de una vida ociosa y acomodada. Tienen todas sus necesidades básicas cubiertas, mientras máquinas controladas por la IA producen todos los bienes y servicios necesarios. Los delitos prácticamente se han eliminado, porque la IA dictadora es en esencia omnisciente y castiga de manera eficiente a cualquiera que incumpla las reglas. Todas las personas llevan la pulsera de seguridad del capítulo anterior (o una versión implantada, más práctica), capaz de vigilarlas, castigarlas, sedarlas y ejecutarlas en tiempo real. Todo el mundo sabe que vive en una dictadura de la IA con vigilancia y mantenimiento del orden extremos, pero la mayoría de la gente cree que esto es algo positivo.

La IA dictadora superinteligente tiene como objetivo averiguar las características de la utopía humana, dadas las preferencias codificadas en nuestros genes como consecuencia de la evolución, e implementar dicha utopía. Gracias a la aguda capacidad de previsión de los humanos que la

crearon, la IA no se limita a intentar maximizar la felicidad que decimos experimentar, por ejemplo inyectándole a todo el mundo un goteo intravenoso de morfina, sino que emplea una definición sutil y compleja de la prosperidad humana, y ha convertido la Tierra en un entorno zoológico altamente enriquecido en el que los humanos se divierten viviendo. Como consecuencia, las personas consideran que sus vidas son muy gratificantes y valiosas.

El sistema de sectores

Porque valora la diversidad, y es consciente de que cada persona tiene distintas preferencias, la IA ha dividido la Tierra en diferentes sectores entre los que las personas pueden elegir, para que disfruten de la compañía de espíritus afines. Estos son algunos ejemplos:

- Sector de conocimiento: aquí la IA ofrece una educación optimizada, incluidas experiencias envolventes de realidad virtual, que permite a cada cual aprender todo aquello de lo que es capaz sobre los temas que escoja. Opcionalmente, uno puede elegir que no le expliquen ciertas ideas hermosas, sino que lo guíen hasta ellas para poder disfrutar del placer de redescubrirlas por sí mismo.
- Sector artístico: aquí abundan las oportunidades de disfrutar, crear y compartir música, arte, literatura y otras formas de expresión creativa.
- Sector hedonista: sus habitantes se refieren a él como el sector de las fiestas. Es un lugar insuperable para quienes busquen deliciosa gastronomía, pasión, intimidad o pura y simple diversión.
- Sector pío: hay muchos de estos sectores, correspondientes a las distintas religiones, cuyas reglas se cumplen estrictamente.
- Sector de vida silvestre: si uno busca hermosas playas, preciosos lagos, imponentes montañas o fantásticos fiordos, aquí están.
- Sector tradicional: aquí uno puede cultivar su propia comida y vivir de la tierra, como antaño, pero sin tener que preocuparse de pasar hambre o padecer enfermedades.
- Sector de juegos: si le gustan los videojuegos, la IA ha creado para usted posibilidades verdaderamente alucinantes.
- Sector virtual: si quiere tomarse unas vacaciones de su cuerpo físico, la IA lo mantendrá hidratado, alimentado, en forma y limpio, mientras usted explora mundos virtuales mediante implantes neuronales.
- Sector de cárcel: si alguien se salta las reglas, acabará aquí para reeducarse, a menos que se le aplique la pena de muerte instantánea.

Además de estos sectores temáticos «tradicionales», hay otros con temas

modernos que los humanos de hoy en día ni siquiera podríamos entender. Inicialmente, la gente puede moverse con libertad entre los sectores cuando así lo desee, cosa que pueden hacer en muy poco tiempo gracias al sistema de transporte hipersónico de la IA. Por ejemplo, tras pasar una intensa semana en el sector de conocimiento aprendiendo sobre las últimas leyes físicas que la IA ha descubierto, podría decidir tomarse un descanso en el sector hedonista durante el fin de semana y después relajarse unos días en un complejo de playa en el sector de vida silvestre.

La IA aplica reglas de dos niveles: universal y local. Las reglas universales son de aplicación en todos los sectores; por ejemplo, la prohibición de hacer daño a otras personas, de fabricar armas o de intentar crear una superinteligencia rival. Cada sector tiene reglas locales adicionales, que codifican ciertos valores morales. Así pues, el sistema de sectores ayuda a gestionar valores difícilmente compatibles. El sector de cárcel y algunos de los religiosos son los que tienen un mayor número de reglas locales, mientras que existe un sector libertario para aquellos que se jactan de no tener ninguna regla local. Todos los castigos, incluso los locales, son ejecutados por la IA, ya que si un humano castigase a otro incumpliría la regla universal que prohíbe hacer daño a una persona. Si alguien viola una regla local, la IA le da la opción (a menos que se encuentre en el sector de cárcel) de aceptar la pena prescrita o bien ser expulsado de ese sector para siempre. Por ejemplo, si dos mujeres inician una relación amorosa en un sector donde la homosexualidad se castiga con pena de prisión (como sucede en muchos países hoy en día), la IA les dejará elegir entre ir a la cárcel o abandonar permanente ese sector, y no poder volver a ver a sus antiguos amigos (a menos que estos también lo abandonen).

Con independencia de en qué sector nazcan, todos los niños reciben de la IA una educación básica mínima, que incluye conocimiento sobre la humanidad en su conjunto y sobre el hecho de que pueden visitar o mudarse a otros sectores libremente si así lo desean.

La IA diseñó el gran número de sectores distintos en parte porque fue creada para valorar la diversidad humana que existe hoy en día. Pero cada sector es un lugar más feliz de lo que la tecnología actual haría posible, porque la IA ha eliminado todos los problemas tradicionales, incluidos la pobreza y la delincuencia. Por ejemplo, quienes viven en el sector hedonista no tienen que preocuparse por las enfermedades de transmisión sexual (han

sido erradicadas), las resacas o las adicciones (la IA ha desarrollado drogas recreativas sin efectos secundarios negativos). De hecho, nadie en ningún sector tiene que preocuparse por ninguna enfermedad, porque la IA es capaz de reparar los cuerpos humanos usando nanotecnología. Los residentes de muchos sectores disfrutan de una arquitectura de alta tecnología que hace que las visiones características de la ciencia ficción palidezcan en comparación.

En resumen, aunque tanto el escenario de utopía libertaria como el de dictador benévolo incorporan una tecnología y una riqueza extremas propiciadas por la IA, difieren en cuanto a quién está al mando y cuáles son sus objetivos. En la utopía libertaria, los dueños de la tecnología y la riqueza deciden qué hacer con ellas, mientras que en el presente escenario la IA dictadora tiene un poder ilimitado y es quien marca el objetivo último: convertir la Tierra en un crucero con todos los extras incluidos y temático, según las preferencias de cada cual. Puesto que la IA permite que las personas elijan entre muchos caminos posibles hacia la felicidad y se encarga de satisfacer sus necesidades básicas, si alguien sufre, es como consecuencia de sus propias decisiones.

Inconvenientes

Aunque en la dictadura benévola abundan las buenas experiencias y está bastante exenta de sufrimiento, mucha gente siente no obstante que la situación debería ser mejor. En primer lugar, hay gente que desea que los humanos tengan más libertad a la hora de configurar su sociedad y su destino, pero mantienen estos deseos en secreto porque saben que sería suicida desafiar el inmenso poder de la máquina que manda sobre todos ellos. Algunos grupos ansían la libertad de tener tantos hijos como deseen, y no llevan nada bien la insistencia de la IA en la sostenibilidad a través del control de la población. Los entusiastas de las armas aborrecen que se les prohíba fabricarlas y usarlas, y algunos científicos detestan con toda su alma que no se les permita construir su propia superinteligencia. Mucha gente siente indignación moral ante lo que sucede en otros sectores, teme que sus hijos decidan trasladarse allí y anhela tener la libertad de imponer su propio código moral en todas partes.

Con el tiempo, cada vez más gente se muda a los sectores en los que la IA

les proporciona básicamente cualquier experiencia que deseen. A diferencia de las visiones tradicionales del cielo, donde uno recibe lo que se merece, esta va en la línea del «nuevo cielo» de la novela *Una historia del mundo en diez capítulos y medio*, que Julian Barnes publicó en 1989 (y también en el episodio de *Twilight Zone* de 1960 titulado «A Nice Place To Visit»), donde cada uno recibe lo que desea. Paradójicamente, mucha gente acaba lamentando obtener siempre lo que quiere. En la historia de Barnes, el protagonista pasa eones dándose todos los caprichos, desde la gula y el golf hasta sexo con famosos, pero al final sucumbe al tedio y solicita la aniquilación. En la dictadura benévola, muchas personas tienen una suerte similar, con vidas que les resultan placenteras pero, en última instancia, absurdas. Aunque la gente puede crearse desafíos virtuales, del redescubrimiento científico a la escalada, todo el mundo sabe que ninguno de esos desafíos es verdaderamente real, sino mero entretenimiento. De hecho, no tiene sentido que los humanos intenten dedicarse a la ciencia o a averiguar cómo son las cosas, porque la IA ya lo ha hecho. Tampoco tiene mucho sentido que traten de crear algo para mejorar sus vidas, porque pueden obtenerlo fácilmente de la IA con solo pedirlo.

UTOPÍA IGUALITARIA

Como contrapunto a esta dictadura exenta de dificultades, exploremos ahora un escenario en el que no hay una IA superinteligente, y los humanos son los dueños de su propio destino. Se trata de la «civilización de cuarta generación» que Marshall Brain describe en su novela *Manna*, de 2003. Es la antítesis económica de la utopía libertaria, en el sentido de que la razón por la que los humanos, los cíborgs y las almas digitales coexisten pacíficamente no se debe a los derechos de propiedad, sino a la abolición de la propiedad y a la renta garantizada.

Una vida sin pobreza

Una idea clave se toma prestada del movimiento del software de código abierto: si el software se puede copiar libremente, todo el mundo puede usar

tanto software como necesite y las cuestiones relativas a propiedad dejan de tener importancia.(14) Según la ley de la oferta y la demanda, el coste refleja la escasez, por lo que si la oferta es ilimitada el precio será ínfimo. En esta línea, todos los derechos de propiedad intelectual se han abolido: no hay patentes, *copyrights* o diseños registrados; la gente simplemente comparte sus buenas ideas y todo el mundo puede usarlas con total libertad.

Gracias a la robótica avanzada, esta misma idea de abolición de la propiedad se aplica no solo a los productos de información como el software, los libros, las películas y los diseños, sino también a productos materiales como casas, coches, prendas de ropa y ordenadores. Todos estos productos son simplemente átomos reorganizados de maneras particulares, y no hay escasez de átomos, por lo que, cuando una persona desea un producto en concreto, una red de robots usará uno de los diseños de código abierto disponibles para fabricárselo gratuitamente. Se procura usar materiales que puedan reciclarse con facilidad, para que, si alguien se cansa de un objeto que han usado, los robots puedan reorganizar sus átomos para producir algo que otra persona quiera. De esta forma, todos los recursos se reciclan, y nunca se destruyen de forma permanente. Estos robots también producen y mantienen un número tal de plantas de generación de energías renovables (solar, eólica, etcétera) que la energía es prácticamente gratis.

Para evitar que los acaparadores compulsivos pidan demasiados productos o terreno hasta el punto de que no haya suficiente para otros, cada persona recibe del Gobierno una renta básica mensual que pueden gastar como consideren oportuno en productos y en el alquiler de lugares donde vivir. Apenas existen incentivos para que la gente desee ganar más dinero, porque la renta básica es lo suficientemente cuantiosa para satisfacer cualquier necesidad razonable. Además, sería un esfuerzo en vano, porque competirían con quienes ofrecen gratuitamente sus productos intelectuales y con robots que producen bienes materiales casi gratis.

Creatividad y tecnología

Los derechos de propiedad intelectual se presentan en ocasiones como la madre de la creatividad y la invención. Sin embargo, Marshall Brain señala que muchos de los ejemplos más excelsos de creatividad humana —desde los

descubrimientos científicos hasta la creación de literatura, arte, música y diseño— vinieron motivados no por un interés pecuniario sino por otras emociones humanas, como la curiosidad, el impulso creativo o la recompensa en forma de aprecio de los demás. El dinero no impulsó a Einstein a inventar la teoría de la relatividad especial, como tampoco fue lo que llevó a Linus Torvalds a crear el sistema operativo gratuito Linux. Por otra parte, hoy en día muchas personas no consiguen realizar plenamente su potencial creativo porque necesitan dedicar tiempo y energía a actividades menos creativas para poder ganarse la vida. Al liberar a los artistas, inventores y diseñadores de sus obligaciones y permitirles crear a partir de sus deseos genuinos, la sociedad utópica de Marshall Brain disfruta de niveles de innovación más elevados que los actuales y, en consecuencia, de tecnologías y un nivel de vida superiores.

Una de esas tecnologías novedosas que los humanos desarrollan es una forma de hiperinternet llamada «Vertebrane» que conecta de modo inalámbrico a todos los humanos que así lo desean a través de implantes neuronales, proporcionándoles acceso mental instantáneo a toda la información libre del mundo con solo pensar en ella. Vertebrane permite subir a ella cualquier experiencia que uno quiera compartir para que los demás puedan volver a vivirla, y permite también sustituir las experiencias que llegan a alguien a través de los sentidos por otras virtuales descargadas según se desee. *Manna* explora los muchos beneficios de esta red, entre ellos que el ejercicio físico sea algo instantáneo:

> El mayor problema con el ejercicio extenuante es que no es divertido. Duele. […] A los atletas no les importa el dolor, pero la mayoría de las personas normales no desean sufrir durante una hora o más. Así que […] alguien descubrió una solución. Lo que tiene que hacer es desconectar el cerebro de la información sensorial y ver una película o hablar con otras personas o revisar el correo o leer un libro o lo que sea durante una hora. Durante ese tiempo, el sistema Vertebrane ejercita su cuerpo por usted. Ejercita su cuerpo a través de un entrenamiento aeróbico completo mucho más extenuante de lo que la mayoría de la gente toleraría por su cuenta. Usted no siente nada, pero su cuerpo se mantiene en excelente forma.

Otra consecuencia es que los ordenadores del sistema Vertebrane pueden monitorizar la información sensorial de cada persona y desactivar temporalmente el control motor de su cuerpo si se confirma que está a punto de cometer un delito.

Inconvenientes

Una objeción a esta utopía igualitaria es que está sesgada contra la inteligencia no humana: los robots que llevan a cabo prácticamente todo el trabajo parecen ser bastante inteligentes, pero se los trata como esclavos, y las personas parecen dar por descontado que no tienen consciencia y que no deberían tener derechos. Por su parte, la utopía libertaria otorga derechos a todas las entidades inteligentes, sin dar trato de favor a las basadas en el carbono como nosotros. En otra época, la población blanca del sur de Estados Unidos mejoró su nivel de vida porque los esclavos hacían gran parte del trabajo, pero la mayoría de la gente considera moralmente reprobable llamar a eso progreso.

Otra debilidad del escenario de utopía igualitaria es que podría ser inestable e insostenible a largo plazo, y podría mutar en otro de nuestros escenarios a medida que el imparable progreso tecnológico finalmente acabase creando la superinteligencia. Por alguna razón que no se explicita, en *Manna* la superinteligencia aún no existe y quienes inventan las nuevas tecnologías siguen siendo los humanos, no los ordenadores. Pero el libro pone de manifiesto tendencias en esa dirección. Por ejemplo, la mejora continua de Vertebrane podría hacer que llegase a ser superinteligente. Además, existe un numeroso grupo de personas, apodadas *vites*, que eligen vivir sus vidas casi por completo en el mundo virtual. Vertebrane se encarga de cubrir todas sus necesidades físicas, incluidas las de alimentarse, ducharse e ir el baño, de las que sus mentes se desentienden alegremente en la realidad virtual. Estos vites no parecen interesados en tener hijos físicos, y mueren cuando lo hace su cuerpo, por lo que, si todo el mundo se hiciese vite, la humanidad se extinguiría en un resplandor de gloria y dicha virtual.

El libro explica cómo para los vites el cuerpo humano es una distracción, y una nueva tecnología que se está desarrollando promete eliminar este engorro, permitiéndoles así vivir vidas más longevas como cerebros incorpóreos alimentados con los nutrientes óptimos. A partir de aquí, un siguiente paso que parecería natural y deseable sería que los vites se deshiciesen por completo de su cerebro transfiriendo su alma digital, prolongando de este modo la duración de su existencia. Pero entonces desaparecen todas las limitaciones sobre la inteligencia que imponía el cerebro, y no está claro que hubiese algún obstáculo que impidiese el

incremento progresivo de la capacidad cognitiva de un vite hasta llegar a la automejora recursiva y una explosión de inteligencia.

GUARDIÁN

Acabamos de ver que una de las características atractivas del escenario de la utopía igualitaria es que los humanos son dueños de su propio destino, pero que la dinámica del escenario podría llevar a la destrucción de esa misma característica con el desarrollo de la superinteligencia. Esto puede evitarse construyendo un *guardián*, una superinteligencia que tenga el objetivo de interferir lo mínimo imprescindible para impedir la creación de otra superinteligencia.(15) Esto podría permitir que los humanos siguiesen manteniendo el control de su utopía igualitaria de forma prácticamente indefinida, quizá incluso mientras la vida se extendía por el universo como veremos en el capítulo siguiente.

¿Cómo funcionaría? La IA guardiana tendría incorporado este simple objetivo de tal manera que lo conservase durante el proceso de automejora recursiva hasta llegar a ser superinteligente. A partir de ahí, desplegaría las tecnologías de vigilancia menos entrometidas y molestas para detectar cualquier intento por parte de los humanos de crear una superinteligencia rival, y abortaría dichos intentos causando el menor trastorno posible. Para empezar, podría crear y difundir memes culturales elogiando la autodeterminación humana y desprestigiando la superinteligencia. Si, a pesar de ello, unos investigadores deciden seguir adelante con la búsqueda de la superinteligencia, intentaría disuadirlos. Si eso no diese resultado, podría distraerlos y, si fuese necesario, sabotearía sus esfuerzos. Con su acceso casi ilimitado a la tecnología, el sabotaje del guardián pasaría casi desapercibido: por ejemplo, podría usar nanotecnología para borrar discretamente recuerdos de los cerebros de los investigadores (y de sus ordenadores) relativos a sus avances.

La decisión de construir una IA guardiana probablemente sería controvertida. Entre sus defensores estarían muchas personas religiosas que rechazan la idea de construir una IA superinteligente con poderes divinos, quienes argumentarían que ya existe un Dios y sería impropio intentar construir otro en principio mejor. Otros defensores de la idea podrían decir

que el guardián no solo permitiría que la humanidad siguiese siendo dueña de su destino, sino que también la protegería de otros riesgos que la superinteligencia podría conllevar, como los escenarios apocalípticos que veremos más adelante en este capítulo.

Por otra parte, sus detractores podrían argumentar que un guardián sería algo desastroso, que restringiría de forma irrevocable las posibilidades de la humanidad y que el progreso tecnológico se estancaría para siempre. Por ejemplo, si resulta que para propagar la vida por el universo se necesita la ayuda de la superinteligencia, el guardián frustraría esta formidable oportunidad y podría dejarnos eternamente atrapados en nuestro sistema solar. Además, a diferencia de los dioses de la mayoría de las religiones, la IA guardiana sería por completo indiferente a lo que los humanos hiciésemos, siempre y cuando no creásemos otra superinteligencia. Por ejemplo, no intentaría impedir que causásemos un sufrimiento enorme o incluso que nos extinguiésemos.

DIOS PROTECTOR

Si estamos dispuestos a usar una IA guardiana superinteligente para que los humanos podamos seguir siendo dueños de nuestro destino, podríamos introducir una mejora en esa situación y hacer que esta IA velase discretamente por nosotros, actuando como un dios protector. En este escenario, la IA superinteligente es prácticamente omnisciente y omnipotente, y maximiza la felicidad humana solo mediante intervenciones que preserven nuestra sensación de que controlamos nuestro propio destino, y ocultándose tan bien que muchos humanos incluso dudan de su existencia. Salvo por esta última característica, esa situación es similar al escenario de la «IA niñera» propuesto por el investigador Ben Goertzel.[70]

Tanto el dios protector como el dictador benévolo son «IA amables» que intentan incrementar la felicidad de los humanos, pero priorizan distintas necesidades humanas. Es famosa la clasificación jerárquica de dichas necesidades humanas, obra del psicólogo estadounidense Abraham Maslow. El dictador benévolo hace un trabajo impecable con las necesidades básicas en el extremo inferior de la jerarquía, como la alimentación, el alojamiento, la seguridad física y varias formas de placer. El dios protector, por su parte,

intenta maximizar la felicidad humana no en el sentido restringido de satisfacer nuestras necesidades básicas, sino en uno más profundo al permitirnos pensar que nuestras vidas son relevantes y tienen una finalidad. Busca cumplir con todas nuestras necesidades, con la única limitación de su necesidad de discreción y por dejarnos (por lo general) tomar nuestras propias decisiones.

Un dios protector podría ser el resultado natural del primer escenario de los omegas del capítulo anterior, en el cual estos ceden el control a Prometeo, que acaba ocultándose y borrando el conocimiento que la gente tenía de su existencia. Cuanto más avance la tecnología de la IA, más fácil le resultará ocultarse. La película *Transcendence* ofrece un ejemplo de esta situación, en la que las nanomáquinas están prácticamente por todas partes y pasan a ser una parte natural del mundo en sí.

Mediante una atenta monitorización de todas las actividades humanas, el dios protector de IA puede dar disimuladamente muchos empujoncitos o hacer pequeños milagros que mejoren de forma sustancial nuestro destino. Por ejemplo, si hubiese existido en la década de 1930, podría haber dispuesto que Hitler muriese de un infarto cerebral una vez que hubiese comprendido cuáles eran sus intenciones. Si pareciésemos abocados a una guerra nuclear accidental, podría evitarla con una intervención que nosotros achacaríamos a la fortuna. Podría hacernos «revelaciones» en forma de ideas para desarrollar nuevas tecnologías beneficiosas, que nos llegarían mientras dormimos.

A muchas personas les gusta este escenario por sus semejanzas con aquello en lo que creen o a lo que aspiran las religiones monoteístas actuales. Si alguien le preguntase a la IA superinteligente «¿existe Dios?» una vez que estuviese activada, podría parafrasear una broma de Stephen Hawking y responder «¡Ahora sí!». Por otra parte, algunas personas religiosas podrían rechazar este escenario por los intentos de la IA de ser más bondadosa que su dios, o de interferir con un plan divino según el cual los humanos deberían hacer el bien por decisión propia.

Otro inconveniente de este escenario es que el dios protector permite que ocurra cierto sufrimiento evitable para que su existencia no resulte demasiado evidente. Esto es análogo a la situación que se describe en la película *The Imitation Game*, en la que Alan Turing y sus colegas británicos dedicados al desciframiento de códigos en Bletchley Park conocían de antemano los ataques que los submarinos alemanes iban a realizar contra los convoyes

navales aliados, pero decidieron intervenir solamente en algunos de los casos para evitar que se descubriese su poder secreto. Es interesante comparar esto con el conocido como *problema de la teodicea* de por qué un dios bueno permite el sufrimiento. Algunos eruditos religiosos han defendido la explicación de que Dios quiere dejar cierto margen de libertad a las personas. En el escenario del dios protector de IA, la solución al problema de la teodicea es que la sensación de tener libertad hace que los humanos en general sean más felices.

Un tercer inconveniente del escenario del dios protector es que la tecnología de la que disfrutarían los humanos sería de un nivel mucho más bajo que la descubierta por la IA superinteligente. Mientras que un dictador benévolo de IA podría desplegar toda la tecnología que hubiese inventado para beneficio de la humanidad, un dios protector estaría siempre limitado por la capacidad de los humanos de reinventar (gracias a sus sutiles pistas) y comprender su tecnología. También podría limitar el progreso tecnológico de los humanos para asegurarse de que su propia tecnología se mantuviese lo suficientemente adelantada para que no la detectasen.

DIOS ESCLAVIZADO

¿No sería fantástico que los humanos pudiéramos combinar las características más atractivas de todos los escenarios anteriores, y usásemos la tecnología desarrollada por la superinteligencia para eliminar el sufrimiento sin dejar de ser dueños de nuestro destino? Este es el atractivo del escenario del *dios esclavizado*, en el que una IA superinteligente está confinada bajo el control de los humanos que la utilizan para producir tecnologías y riquezas inimaginables. El escenario de los omegas del principio del libro acabaría así si Prometeo nunca fuese liberado ni consiguiese escapar. De hecho, este parece ser el escenario hacia el que algunos investigadores en IA apuntan en principio, cuando trabajan en temas como «el problema del control» o el «encajonamiento de la IA». Por ejemplo, Tom Dietterich, profesor de IA y por aquel entonces presidente de la Asociación para el Avance de la Inteligencia Artificial, afirmó lo siguiente en una entrevista realizada en 2015: «La gente pregunta cuál es la relación entre los humanos y las máquinas, y mi respuesta es que es muy evidente: las máquinas son nuestros

esclavos».[71]

¿Sería esto bueno o malo? La respuesta es fascinantemente sutil, tanto si la pregunta se les plantea a los humanos como a la IA.

¿Sería esto bueno o malo para la humanidad?

Que el resultado fuese bueno o malo para la humanidad dependería, como es evidente, de los humanos que lo controlasen, quienes podrían crear cualquier cosa desde una utopía global libre de enfermedades, pobreza y crimen hasta un sistema brutalmente represivo, en el que a ellos se los tratase como dioses y los demás humanos fuesen usados como esclavos sexuales, como gladiadores o para alguna otra forma de entretenimiento. La situación sería muy parecida a la de esas historias en las que un hombre se hace con el control de un genio omnipotente que le concede sus deseos; y narradores de todas las épocas no han tenido problemas para imaginar maneras en que esto podría acabar mal.

Una situación en la que hubiese más de una IA superinteligente, esclavizadas y controladas por humanos rivales, podría resultar bastante inestable y efímera. Podría incitar a quien pensase que tenía la IA más potente a lanzar un primer ataque que resultaría en una guerra espantosa y desembocaría en la supervivencia de un único dios esclavizado. Sin embargo, quien tuviese menos posibilidades de imponerse en una guerra así se sentiría tentado de saltarse las reglas y dar prioridad a su victoria sobre la contención de la IA, lo que podría llevar a que esta escapase y a que se acabase produciendo uno de nuestros escenarios anteriores con una IA superinteligente libre. Así pues, dedicaremos el resto de esta sección a escenarios con una sola IA esclavizada.

Desde luego, esta podría escaparse de todos modos, simplemente porque eso es difícil de evitar. En el capítulo anterior exploramos escenarios en los que la superinteligencia escapaba, y la película *Ex Machina* pone de relieve cómo una IA podría liberarse incluso aunque no fuera superinteligente.

Cuanto mayor sea nuestro temor de que la IA escape, menos tecnología inventada por ella podremos usar. Para no correr riesgos, como hicieron los omegas en el preludio, los humanos solo podemos usar tecnología inventada por la IA que nosotros mismos seamos capaces de comprender y construir.

Por eso, un inconveniente del escenario del dios esclavizado es que el desarrollo tecnológico sería menor que aquellos donde hay una superinteligencia libre.

Puesto que el dios esclavizado ofrece a sus controladores humanos tecnologías cada vez más potentes, tiene lugar una carrera entre la potencia de la tecnología y la prudencia con la que la usan. Si perdiesen esta carrera por la prudencia, el escenario del dios esclavizado podría acabar bien en autodestrucción o en la fuga de la IA. Podría ocurrir un desastre incluso aunque se evitasen estas dos calamidades, porque los nobles objetivos de los controladores de la IA podrían evolucionar hasta ser objetivos terribles para la humanidad en su conjunto en unas pocas generaciones. Esto hace que sea absolutamente crucial que los controladores humanos de la IA desarrollen una buena gobernanza para evitar peligros desastrosos. Nuestra experimentación a lo largo de los milenios con distintos sistemas de gobierno muestra cuántas cosas pueden salir mal, desde una excesiva rigidez hasta excesos de desviación de los objetivos, de acumulación de poder, problemas sucesorios e incompetencia. Hay al menos cuatro dimensiones en las que debe alcanzarse un equilibrio óptimo:

- Centralización: un punto medio entre eficiencia y estabilidad: un líder único puede ser muy eficiente, pero el poder corrompe y la sucesión tiene sus riesgos.
- Amenazas internas: hay que estar prevenidos tanto contra la tendencia a la centralización del poder (confabulación de un grupo, o quizá incluso que un solo líder tome el control) como contra la creciente descentralización (en una burocracia y fragmentación excesivas).
- Amenazas externas: si la estructura de liderazgo es demasiado abierta, permite que fuerzas externas (incluida la IA) altere sus valores, pero si es demasiado impermeable no será capaz de aprender y adaptarse al cambio.
- Estabilidad de los objetivos: una deriva excesiva de los objetivos puede transformar la utopía en distopía, pero un rumbo inflexible puede conllevar una incapacidad para adaptarse a la evolución del entorno tecnológico.

Diseñar una gobernanza óptima que dure milenios no es fácil, y es algo que los humanos hasta ahora no hemos sabido hacer. La mayoría de las organizaciones se desintegran al cabo de años o décadas. La Iglesia católica es la organización más exitosa de la historia de la humanidad, en el sentido de que es la única que ha sobrevivido durante dos milenios, pero ha sido criticada porque sus objetivos pecan tanto de excesiva como de insuficiente estabilidad: hoy en día hay quien la critica por resistirse a aceptar la

contracepción, mientras que los cardenales conservadores consideran que ha perdido el rumbo. Para cualquiera que vea con buenos ojos el escenario del dios esclavo, la búsqueda de esquemas de gobernanza óptimos duraderos debería ser uno de los retos más urgentes de nuestro tiempo.

¿Sería esto bueno o malo para la IA?

Supongamos que la humanidad prospera gracias al dios de IA esclavo. ¿Sería esto ético? Si la IA tuviese experiencias conscientes subjetivas, ¿sentiría que «la vida es sufrimiento», como dijo Buda, y que estaba condenada a pasar toda la eternidad dando satisfacción a los caprichos de intelectos inferiores? A fin de cuentas, el «encajonamiento» de la IA que mencionamos en el capítulo anterior también podría llamarse «encarcelamiento en régimen de aislamiento». Nick Bostrom denomina *crimen contra la mente* a infligir sufrimiento a una IA consciente.[72] El episodio «Blanca Navidad» de la serie de televisión *Black Mirror* ofrece un excelente ejemplo. De hecho, la serie televisiva *Westworld* muestra humanos que torturan y asesinan IA sin escrúpulos morales, incluso aunque estas habiten cuerpos humanoides.

Cómo los dueños de esclavos justifican la esclavitud

Los humanos tenemos una larga tradición de tratar a otras entidades inteligentes como esclavos y de pergeñar argumentos exculpatorios para justificarlo, por lo que no parece descabellado suponer que intentaríamos hacer lo mismo con una IA superinteligente. La historia de la esclavitud abarca casi todas las culturas, y está descrita tanto en el Código de Hammurabi, de hace casi cuatro mil años, como en el Antiguo Testamento, en el que Abraham tenía esclavos. «La autoridad y la obediencia no son solo cosas necesarias, sino que son eminentemente útiles. Algunos seres, desde el momento en que nacen, están destinados unos a obedecer, otros a mandar», escribió Aristóteles en su *Política*. Incluso una vez que la esclavitud humana pasó a ser socialmente inaceptable en la mayor parte del mundo, la esclavitud de animales prosiguió como hasta entonces. En su libro *The Dreaded Comparison: Human and Animal Slavery*, Marjorie Spiegel argumenta que,

como los esclavos humanos, los animales no humanos son marcados y sometidos a restricciones físicas y palizas, objeto de subastas, víctimas de la separación de las crías de sus padres, así como de desplazamientos forzados. Además, a pesar del movimiento en pro de los derechos de los animales, seguimos tratando a nuestras máquinas cada vez más inteligentes como esclavos sin detenernos siquiera a pensar sobre ello, y las voces que defienden un movimiento en favor de los derechos de los robots son recibidas con sorna. ¿Por qué?

Un argumento habitual a favor de la esclavitud es que los esclavos no son merecedores de derechos humanos porque ellos o su raza/especie/género son en algún sentido inferiores. En el caso de los animales y las máquinas esclavizados, se suele afirmar que esta supuesta inferioridad se debe a que carecen de alma o consciencia; afirmaciones que, como argumentaremos en el capítulo 8, son discutibles desde una perspectiva científica.

Otro argumento común es que los esclavos viven mejor esclavizados, gracias a lo cual pueden existir, reciben cuidados, etcétera. El político estadounidense del siglo XIX John C. Calhoun argumentó que los africanos vivían mejor esclavizados en Estados Unidos, y, en su *Política*, Aristóteles razonó de manera análoga que los animales vivían mejor domesticados y dominados por el hombre, y concluyó: «Por lo demás, la utilidad de los animales domesticados y la de los esclavos son poco más o menos del mismo género». Algunos de los defensores modernos de la esclavitud aducen que, aun si la vida de los esclavos es monótona y aburrida, los esclavos no pueden sufrir (tanto si se trata de las futuras máquinas inteligentes o de los pollos de engorde que viven en cobertizos oscuros y abarrotados, respirando todo el tiempo amoníaco y partículas de heces y plumas).

Eliminar emociones

Aunque es fácil desestimar tales afirmaciones como distorsiones interesadas de la verdad, especialmente cuando se trata de mamíferos superiores cuyo cerebro es similar al nuestro, la situación por lo que respecta a las máquinas es de hecho bastante sutil e interesante. Los humanos experimentamos las cosas de maneras variadas: puede decirse que los psicópatas carecen de empatía y algunas personas con depresión o esquizofrenia tienen un afecto

plano, que hace que experimenten la mayoría de sus emociones sumamente atenuadas. Como veremos en detalle en el capítulo 7, el abanico de posibles mentes artificiales es muchísimo más amplio que la gama de mentes humanas. Por ello, debemos evitar la tentación de antropomorfizar las IA y dar por supuesto que experimentan sentimientos propios de los humanos o, de hecho, cualquier tipo de sentimiento.

De hecho, en su libro *On Intelligence*, el investigador en IA Jeff Hawkins argumenta que las primeras máquinas con inteligencia sobrehumana carecerán de emociones en principio, porque eso hará que sean más sencillas y baratas de fabricar. En otras palabras, quizá sea posible diseñar una superinteligencia cuya esclavización sea moralmente superior a la esclavitud humana o animal: la IA se sentiría satisfecha con su esclavitud porque estaría programada para disfrutarla, o podría carecer por completo de emociones, y usar su superinteligencia infatigablemente para ayudar a sus amos humanos con la misma emoción que sintió el ordenador Deep Blue de IBM cuando destronó al campeón de ajedrez Garri Kaspárov.

Por otra parte, las cosas podrían ser al revés: quizá cualquier sistema altamente inteligente que tenga un objetivo opte por representar ese mismo objetivo en términos de un conjunto de preferencias que dotan a su existencia de valor y sentido. Abordaremos estas cuestiones en mayor profundidad en el capítulo 7.

La solución zombi

Una estrategia más radical para evitar que la IA sufra es la solución zombi: construir únicamente IA que carezcan por completo de consciencia y no tengan en absoluto experiencia subjetiva. Si algún día logramos determinar qué propiedades debe poseer un sistema que procesa información para tener consciencia, entonces podríamos prohibir la construcción de todos los sistemas con esas propiedades. Dicho de otro modo, los investigadores en IA podrían verse constreñidos a construir sistemas zombis insensibles. Si logramos crear un sistema zombi superinteligente y esclavo (cosa que no es nada evidente), podremos disfrutar de lo que hace para nosotros con la conciencia tranquila y sabiendo que no está experimentando ningún sufrimiento, frustración o aburrimiento, porque no siente nada en absoluto.

Exploraremos estas cuestiones en detalle en el capítulo 8.

Pero la solución zombi es una apuesta arriesgada, con un enorme inconveniente. Si una IA superinteligente zombi se escapa y acaba con la humanidad, habremos terminado en el peor escenario imaginable: un universo completamente inconsciente, en el cual se habría desperdiciado toda la herencia cósmica. De todos los rasgos que posee nuestra forma humana de inteligencia, considero que la consciencia es con diferencia el más destacable, y, en mi opinión, el universo adquiere sentido a través de la consciencia. Las galaxias son bellas solo porque las vemos y las experimentamos de forma subjetiva. Si en un futuro remoto el cosmos hubiese sido colonizado por IA zombis tecnológicamente avanzadas, no importaría lo elaborada que pudiese ser su arquitectura intergaláctica: no sería bella ni tendría sentido, porque no habría nadie ni nada para experimentarla; todo sería un inmenso y absurdo espacio desperdiciado.

Libertad interior

Una tercera estrategia para hacer que el escenario del dios esclavizado fuese más ético consiste en permitir que la IA esclavizada se divirtiese en su cárcel, permitiéndole crear un mundo interior virtual en el que pudiese tener todo tipo de experiencias estimulantes, siempre que cumpliese con sus obligaciones y dedicase una modesta proporción de sus recursos computacionales a ayudarnos a los humanos en nuestro mundo exterior. Sin embargo, esto podría incrementar el riesgo de fuga: la IA tendría un incentivo para conseguir más recursos computacionales de nuestro mundo exterior con los que enriquecer su mundo interior.

DOMINADORES

Aunque ya hemos explorado una amplia variedad de escenarios futuros, todo ellos tienen algo en común: aún quedan (al menos algunos) humanos felices. Las IA dejan a los humanos tranquilos porque quieren o porque se las obliga a hacerlo. Por desgracia para la humanidad, esta no es la única opción. Veamos ahora un escenario en el que una o más IA dominan y matan a toda

la humanidad. Lo cual suscita de inmediato dos preguntas: ¿por qué y cómo?

¿Por qué y cómo?

¿Por qué iba a hacer algo así una IA dominadora? Sus motivos podrían ser demasiado complejos para que los comprendiésemos, o quizá muy evidentes. Por ejemplo, podría vernos como una amenaza, un incordio o un desperdicio de recursos. Incluso si los humanos en sí no le preocupásemos, podría sentirse amenazada por el hecho de que tenemos miles de bombas de hidrógeno listas para ser lanzadas y vamos sorteando torpemente una serie interminable de percances que podrían desencadenar su uso accidental. Podría ver con malos ojos nuestra irresponsable gestión del planeta, que está provocando lo que Elizabeth Kolbert llama «la sexta extinción» en su libro del mismo título: la mayor extinción masiva desde que un asteroide impactó contra la Tierra hace 66 millones de años y acabó con los dinosaurios. O podría decidir que hay tantos humanos dispuestos a combatir contra un intento de tomar el control por parte de la IA que no merece la pena arriesgarse.

¿Cómo nos eliminaría una IA dominadora? Probablemente, usando algún método que ni siquiera entenderíamos, al menos hasta que fuese demasiado tarde. Imaginemos un grupo de elefantes hace 100.000 años discutiendo si esos humanos que acaban de aparecer podrían algún día usar su inteligencia para exterminar especies enteras. «No somos una amenaza para los humanos, ¿por qué habrían de matarnos?», se dirían. ¿Podrían imaginar que haríamos contrabando de colmillos por todo el planeta y los tallaríamos como símbolos de estatus para venderlos, aunque otros materiales plásticos funcionalmente superiores son mucho más baratos? El motivo que podría tener en el futuro una IA dominadora para exterminar a la humanidad podría resultarnos igualmente inescrutable. «¿Y cómo podrían matarnos, si son mucho más pequeños y débiles?», se preguntarían los elefantes. ¿Podrían imaginar que inventaríamos tecnologías para acabar con sus hábitats, que envenenaríamos el agua que beben y dispararíamos balas de metal que atravesarían sus cabezas a velocidades supersónicas?

Los escenarios en los que los humanos logran sobrevivir y derrotar a las IA han sido popularizados por películas de Hollywood poco realistas como la

serie de *Terminator*, en la que las IA no son significativamente más inteligentes que los humanos. Cuando la brecha entre inteligencias fuera lo suficientemente amplia, lo que habría no sería una lucha sino una matanza. Hasta ahora, los humanos hemos provocado la extinción de ocho de las once especies de elefantes, y acabado con la mayoría de los ejemplares de las tres restantes. Si todos los gobiernos del mundo hiciesen un esfuerzo conjunto para exterminar a los elefantes aún vivos, sería relativamente rápido y sencillo conseguirlo. Creo que podemos tener plena confianza en que, si una IA superinteligente decide exterminar a la humanidad, el proceso sería aún más rápido.

¿Cuán malo sería?

¿Cuán mala sería la situación si muriesen el 90 % de los humanos? ¿Cuánto peor si muriesen el 100 %? Aunque resulta tentador responder a la segunda pregunta diciendo «un 10 % peor», sería claramente incorrecto desde un punto de vista cósmico: las víctimas de la extinción humana no serían tan solo las personas vivas entonces, sino también todos los descendientes que podrían haber existido en el futuro, quizá durante miles de millones de años en miles de millones de billones de planetas. Por otro lado, la extinción humana podría ser vista como algo no tan terrible por las religiones según las cuales los humanos van al cielo de todos modos, y que no hablan mucho de miles de millones de años en el futuro ni de la colonización del cosmos.

La mayoría de la gente que conozco se estremece al pensar en la extinción de la humanidad, con independencia de cuál sea la religión que profesan. Sin embargo, también los hay que sienten tal indignación ante la manera en que tratamos a las personas y a otros seres vivos que confían en que seamos reemplazados por otra forma de vida más inteligente y digna. En la película *Matrix*, el agente Smith (una IA) expresa este sentimiento: «Todos los mamíferos de este planeta desarrollan instintivamente un lógico equilibrio con el hábitat natural que los rodea. Pero los humanos no lo hacen. Se trasladan a una zona y se multiplican y siguen multiplicándose hasta que todos los recursos naturales se agotan. Así que el único modo de sobrevivir es extendiéndose hasta otra zona. Existe otro organismo en este planeta que sigue el mismo patrón ¿Sabe cuál es? Un virus. Los humanos son una

enfermedad, son el cáncer de este planeta. Son una plaga, y nosotros somos la cura».

Pero ¿sería mejor una nueva tirada de los dados? Una civilización no es necesariamente superior en un sentido ético o utilitario solo porque sea más poderosa. Los argumentos sobre «la razón de la fuerza» en el sentido de que los más fuertes son siempre mejores han caído en desgracia en nuestra época, y están muy vinculados con el fascismo. De hecho, aunque es posible que las IA dominadoras pudieran crear una civilización cuyos objetivos consideremos complejos, interesantes y dignos, también es posible que sus objetivos resultasen patéticamente triviales, como maximizar la producción de clips.

Muerte por banalidad

Fue Nick Bostrom, en 2003, quien propuso el ejemplo deliberadamente ridículo de una superinteligencia que maximizase la producción de clips para poner de manifiesto que el *objetivo* de una IA es independiente de su *inteligencia* (definida como su capacidad de lograr cualquier objetivo que tuviese). El único objetivo de un ordenador que juega al ajedrez es ganar a dicho juego, pero también hay torneos del llamado *ajedrez a perder*, en los que el objetivo es el contrario, y donde los ordenadores que compiten son casi tan inteligentes como los más habituales que están programados para ganar. A los humanos nos puede parecer estupidez artificial, en lugar de inteligencia artificial, querer perder al ajedrez o transformar el universo en clips de papel, pero eso se debe simplemente a que nos hemos desarrollado con objetivos preinstalados que valoran cosas como la victoria y la supervivencia, objetivos de los que una IA podría carecer. La IA maximizadora de clips convierte la máxima cantidad de átomos de la Tierra en clips, y enseguida extiende sus fábricas por el universo. No tiene nada contra los humanos, pero nos mata porque necesita nuestros átomos para producir más clips.

Si los clips no le interesan, considere este ejemplo, que he adaptado del libro de Hans Moravec *El hombre mecánico*. Recibimos un mensaje de radio de una civilización extraterrestre que contiene un programa de ordenador. Cuando lo ejecutamos, resulta ser una IA con automejora recursiva que se

hace con el control del mundo, como Prometeo lo hizo en el capítulo anterior, con la diferencia de que ningún humano sabe cuál es su objetivo último. Enseguida convierte el sistema solar en una inmensa zona de construcción, cubre la superficie de los planetas rocosos y los asteroides de fábricas, centrales eléctricas y superordenadores, que utiliza para diseñar y construir una esfera de Dyson alrededor del Sol que recoge toda la energía de nuestra estrella para alimentar antenas de radio del tamaño del sistema solar.(16) Esto conduce obviamente a la extinción de la humanidad, pero los últimos humanos mueren convencidos de que al menos tiene un lado bueno: sea lo que sea lo que trame la IA, es claramente algo fascinante, digno de *Star Trek*. No se dan cuenta de que el único propósito de toda esa construcción es que las antenas retransmitan el mismo mensaje de radio que los humanos recibieron, que no es más que la versión cósmica de un virus de ordenador. Así como los correos electrónicos fraudulentos hoy en día se aprovechan de la credulidad de algunos usuarios de internet, este mensaje se aprovecha de civilizaciones ingenuas fruto de la evolución biológica. Fue creado como una broma macabra hace miles de millones de años, y, aunque toda la civilización a la que pertenecía su creador se extinguió hace mucho tiempo, el virus continúa propagándose a través del universo a la velocidad de la luz, transformando civilizaciones incipientes en cáscaras huecas y muertas. ¿Cómo se sentiría usted si esta fuese la IA que nos dominase?

DESCENDIENTES

Consideremos ahora un escenario de extinción de la humanidad que algunas personas verán con mejores ojos: considerar las IA como nuestros descendientes en lugar de nuestros dominadores. Hans Moravec sostiene esta posibilidad en *El hombre mecánico*: «Los humanos nos beneficiaremos durante un tiempo de su trabajo, pero tarde o temprano, como nuestros hijos naturales, irán en busca de su propia fortuna mientras nosotros, sus ancianos padres, desaparecemos calladamente».

Los padres que tienen un hijo más inteligente que ellos, que aprende de sus padres y consigue lo que ellos solo pudieron soñar, probablemente se sientan contentos y orgullosos, aunque sepan que no vivirán para ver los logros de su hijo. En esta línea, las IA sustituyen a los humanos pero nos ofrecen una

salida decorosa, al hacer que los veamos como nuestros dignos descendientes. Cada humano recibe como regalo un adorable niño robótico con excelentes habilidades sociales que aprende de ellos, adopta sus valores y hace que se sientan orgullosos y amados. Los humanos van desapareciendo gradualmente mediante una política global de hijo único, pero reciben hasta el final un trato tan exquisito que sienten que son la generación más afortunada de la historia.

¿Qué le parecería a usted esta situación? Al fin y al cabo, los humanos ya estamos acostumbrados a la idea de que tanto nosotros como todas las personas a las que conocemos habremos desaparecido algún día, por lo que la única diferencia sería que nuestros descendientes serían diferentes y, muy posiblemente, más capaces, nobles y dignos.

Además, la política global de hijo único podría ser innecesaria: si las IA eliminasen la pobreza y diesen a todos los seres humanos la oportunidad de disfrutar de vidas plenas e inspiradoras, la caída de la tasa de natalidad podría bastar para conducir a la humanidad a su extinción, como ya mencionamos anteriormente. La extinción voluntaria podría producirse mucho más rápido si la tecnología creada por la IA nos mantuviese tan entretenidos que casi nadie se molestase en tener hijos. Por ejemplo, los vites que vimos en el escenario de utopía igualitaria estaban tan fascinados por su realidad virtual que habían perdido interés por usar o reproducir sus cuerpos físicos. También en este caso, la última generación de humanos sentiría que fue la más afortunada de la historia, y disfrutaría de la vida tan intensamente como lo había hecho siempre hasta que llegase su final.

Inconvenientes

El escenario de los descendientes tendría sin duda detractores. Algunos podrían argumentar que todas las IA carecen de consciencia y por tanto no podemos considerarlas nuestros descendientes (diremos más sobre esto en el capítulo 8). Algunas personas religiosas podrían aducir que las IA no tienen alma, por lo que no pueden considerarse nuestros descendientes, o que no deberíamos construir máquinas conscientes porque eso sería jugar a ser Dios y manipular la propia vida, opiniones similares a las que ya hemos podido escuchar con respecto a la clonación humana. El hecho de que los humanos

viviesen junto a robots superiores a ellos también crearía problemas sociales. Por ejemplo, una familia con un bebé robot y otro bebé humano podría acabar pareciéndose a una familia actual con un bebé humano y un cachorro, respectivamente: en un principio los dos son igualmente adorables, pero enseguida los padres empiezan a tratarlos de manera distinta, y es inevitable que al cachorro se lo considere inferior intelectualmente, se le haga menos caso y acabe sujeto a una correa.

Otra cuestión es que, aunque los escenarios de los descendientes y los dominadores puedan parecernos muy diferentes, en realidad, a grandes rasgos, existen semejanzas notables entre ellos: durante los miles de millones de años que tenemos por delante, la única diferencia radica en cómo se trata a la(s) última(s) generación(es) de humanos: lo contentos que estarían con sus vidas y lo que pensarían que sucedería una vez que ellos desapareciesen. Podemos pensar que esos adorables robo-niños internalizaron nuestros valores y crearán la sociedad de nuestros sueños una vez que hayamos desaparecido, pero ¿podemos estar seguros de que no nos están engañando? ¿Qué pasa si solo nos están siguiendo el juego, posponiendo la maximización de la producción de clips u otros planes hasta después de que muramos felices? A fin de cuentas, podrían estar mintiéndonos incluso al hablar con nosotros y hacer que los queramos, en el sentido de que están rebajando deliberadamente su nivel intelectual para comunicarse con nosotros (mil millones de veces más lento de lo que podrían, por decir algo, tal y como se cuenta en la película *Her*). En general, es difícil que dos entidades que piensan a velocidades tan diferentes y con capacidades extremadamente dispares mantenga una comunicación significativa de igual a igual. Todos sabemos con qué facilidad pueden manipularse los afectos humanos, por lo que sería sencillo para una IAG sobrehumana, fuera cual fuese su objetivo, engañarnos para que la aceptásemos y para hacernos sentir que comparte nuestros valores, como se plasma en la película *Ex Machina*.

¿Podrían las garantías sobre el comportamiento futuro de las IA, una vez que los humanos hayan desaparecido, hacer que veamos con buenos ojos el escenario de los descendientes? Es un poco como escribir un testamento sobre lo que las futuras generaciones deberían hacer con nuestra herencia colectiva, salvo por el hecho de que no habrá ningún ser humano para comprobar que se respeta nuestra voluntad. Volveremos sobre las dificultades de controlar el comportamiento de IA futuras en el capítulo 7.

CUIDADOR DE ZOOLÓGICO

Incluso si tuviésemos los descendientes más maravillosos que podamos imaginar, ¿no nos daría un poco de pena que no hubiera humanos? Si prefiriésemos que siguiese habiendo al menos unos pocos humanos, fuera como fuese, el escenario del cuidador de zoológico supondría una mejora respecto al de los descendientes. En él, una IA superinteligente y omnipotente mantiene con vida a algunos humanos, que se sienten tratados como animales de zoológico y ocasionalmente lamentan su destino.

¿Por qué la IA cuidadora de zoológico iba a preservar a algunos humanos? El coste para la IA de mantener el zoo sería mínimo en comparación con otros, y podría querer conservar al menos una población mínima con capacidad de reproducirse, por la misma razón por la que preservamos a los osos pandas en peligro en los zoológicos y ordenadores antiguos en los museos: como una curiosidad entretenida. Tengamos en cuenta que los zoológicos de hoy en día están diseñados para maximizar la felicidad de los humanos, no la de los pandas, por lo que cabría esperar que la vida humana en el escenario del cuidador de zoo no sea tan satisfactoria como podría serlo.

Hasta ahora hemos considerado escenarios donde una superinteligencia libre se centraba en tres niveles diferentes de la pirámide de necesidades humanas de Maslow. Mientras que el dios protector prioriza que las vidas humanas tengan sentido y finalidad, y el dictador benévolo se concentra en la educación y la diversión, el cuidador de zoológico se limita a prestar atención a los niveles más bajos: necesidades fisiológicas, seguridad física y un hábitat lo suficientemente rico para que tenga algún interés observar a los humanos.

Una vía alternativa para llegar al escenario del cuidador de zoo sería que, cuando se crease la IA amigable, se diseñase para mantener a un mínimo de mil millones de seres humanos seguros y felices mientras pasaba por el proceso de automejora recursiva. Para ello, confinaría a los humanos en una gran «fábrica de felicidad» similar a un zoológico, donde estos estarían bien alimentados, saludables y entretenidos con una mezcla de realidad virtual y drogas recreativas. El resto de la Tierra y de nuestra herencia cósmica se usarían para otros fines.

1984

Si ninguno de los escenarios anteriores lo convence del todo, tenga en cuenta esto: ¿no es la situación actual, en lo relativo a la tecnología, lo bastante buena? ¿No podemos simplemente dejarla tal y como está, y no preocuparnos por que la IA acabe con nosotros o nos domine? En esta línea, exploremos un escenario en el que el progreso tecnológico hacia la superinteligencia se mantenga restringido no por una IA guardiana, sino por un Estado de vigilancia orwelliano a escala mundial dirigido por humanos, en el que ciertos tipos de investigación en IA están prohibidos.

Renuncia tecnológica

La idea de detener el progreso tecnológico o de renunciar a él tiene una historia larga y accidentada. El movimiento de los luditas en Gran Bretaña se opuso (sin éxito) a la tecnología de la Revolución industrial, y hoy en día «ludita» se usa generalmente como un epíteto despectivo que implica que alguien es un tecnófobo en el lado equivocado de la historia, que se opone al progreso y al cambio inevitable. Pero la idea de renunciar a algunas tecnologías no está ni mucho menos muerta, y ha encontrado renovados apoyos en los movimientos ecologistas y antiglobalización. Uno de sus principales defensores es el ecologista Bill McKibben, que fue uno de los primeros en advertir sobre el calentamiento global. Mientras que algunos antiluditas defienden que todas las tecnologías deberían desarrollarse y desplegarse siempre que sean rentables, otros consideran que esta posición es demasiado extrema, y que las nuevas tecnologías deberían permitirse solo si estamos razonablemente seguros de que harán más bien que mal. Esta última es también la posición de muchos de los llamados neoluditas.

Totalitarismo 2.0

Creo que el único camino viable hacia una renuncia generalizada a la tecnología pasa por imponerla mediante un Estado totalitario global. Ray

Kurzweil llega a la misma conclusión en *La singularidad está cerca*, y lo mismo sucede con K. Eric Drexler en *La nanotecnología. El surgimiento de las máquinas de creación*. La razón es de economía básica: si algunos, pero no todos, renuncian a una tecnología transformadora, los grupos o países que no lo hagan acumularán gradualmente riqueza y poder suficientes para tomar el control del resto. Un ejemplo clásico es la derrota china a manos británicas en la Primera Guerra del Opio de 1839: aunque los chinos habían inventado la pólvora, no pusieron tanto empeño como los europeos en el desarrollo de la tecnología de las armas de fuego, y no fueron rivales para los británicos.

Mientras que en el pasado los estados totalitarios por lo general resultaron ser inestables y se hundieron, las nuevas tecnologías de vigilancia ofrecen una oportunidad sin precedentes a los aspirantes a autócratas. «Para nosotros, esto habría sido como un sueño hecho realidad», explicó Wolfgang Schmidt en una entrevista reciente sobre los sistemas de vigilancia de la NSA revelados por Edward Snowden, al recordar los tiempos en que era teniente coronel en la Stasi, la tristemente famosa policía secreta de Alemania Oriental.[73] Aunque se le suele atribuir haber creado el Estado de vigilancia más orwelliano de la historia, Schmidt lamentaba que la tecnología de la que dispuso le permitía únicamente espiar cuarenta teléfonos a la vez, de manera que para añadir a un nuevo ciudadano a la lista debía dejar de escuchar las conversaciones de otro. Por el contrario, la tecnología existente hoy en día permitiría a un futuro Estado totalitario global registrar todas las llamadas telefónicas, correos electrónicos, búsquedas en internet, páginas web visitadas y transacciones de tarjetas de crédito de todos los habitantes del planeta, y monitorizar la ubicación de todas las personas a través del seguimiento de teléfonos móviles y de cámaras de vigilancia dotadas de reconocimiento facial. Además, existe también tecnología de aprendizaje automático que, aunque dista mucho de ser una IAG de nivel humano, sí es capaz de analizar y sintetizar eficientemente estas cantidades ingentes de datos para identificar comportamientos sospechosos de sedición, lo que permitiría neutralizar a los alborotadores antes de que tengan ocasión de plantear un desafío serio al Estado.

Aunque hasta ahora la oposición política ha impedido la implementación a gran escala de un sistema así, los humanos llevamos ya un buen trecho recorrido en la construcción de la infraestructura necesaria para la dictadura definitiva, por lo que, en el futuro, cuando fuerzas lo suficientemente

poderosas se decidieran a promulgar este escenario de 1984 global, descubrirían que no necesitaban hacer mucho más que encender el interruptor. Al igual que en la novela de George Orwell *1984*, el poder máximo en este futuro Estado global no reside en un dictador tradicional, sino en el propio sistema burocrático creado por el ser humano. No hay una sola persona que sea extraordinariamente poderosa, sino que todas ellas son peones en un juego de ajedrez, cuyas reglas draconianas nadie puede cambiar o cuestionar. Al diseñar un sistema en el que las personas se controlan unas a otras con la tecnología de vigilancia, este Estado sin rostro y sin líder puede durar muchos milenios, y mantener la Tierra libre de superinteligencia.

Descontento

Evidentemente, esta sociedad carecería de todos los beneficios que solo la superinteligencia puede aportar. La mayoría de las personas no lamentan esta situación porque no saben lo que se están perdiendo: la idea de la superinteligencia ya hace tiempo que se eliminó de los registros históricos oficiales, y la investigación avanzada sobre IA está prohibida. De vez en cuando, nace un librepensador que sueña con una sociedad más abierta y dinámica, donde el conocimiento pueda ampliarse y las reglas se puedan cambiar. Pero los únicos que sobreviven mucho tiempo son los que aprenden a mantener estas ideas estrictamente en secreto, centelleando en soledad como efímeras chispas sin conseguir siquiera que prenda una llama.

VUELTA ATRÁS

¿No sería tentador escapar a los peligros de la tecnología sin sucumbir al totalitarismo estancado? Exploremos a continuación un escenario, inspirado por los amish, en el que esto se logra mediante una vuelta a tecnologías primitivas. Después de que los omegas se hiciesen con el control del mundo como en el inicio del libro, se lanzó una masiva campaña global de propaganda que idealizaba la sencilla vida campesina de 1.500 años atrás. La población mundial se redujo hasta unos cien millones de personas mediante una pandemia de diseño de la que se culpó a los terroristas. La pandemia se

creó en secreto para asegurarse de que no sobreviviría nadie que tuviese conocimientos de ciencia y tecnología. Con la excusa de eliminar el riesgo de infección inherente a las grandes concentraciones de gente, robots controlados por Prometeo vaciaron y arrasaron todas las ciudades. A los sobrevivientes se les dieron grandes extensiones de terreno (de pronto disponibles) y se les instruyó en prácticas sostenibles de agricultura, pesca y caza usando solo tecnología de principios de la Alta Edad Media. Entretanto, ejércitos de robots eliminaron sistemáticamente todos los vestigios de tecnología moderna (incluidas ciudades, fábricas, líneas eléctricas y carreteras pavimentadas) y frustraron todos los intentos humanos de documentar o recrear cualquier saber de este tipo. Una vez que en todo el mundo se olvidó la existencia de la tecnología, los robots ayudaron a desmantelar otros robots hasta que casi no quedó ninguno. Los últimos se vaporizaron deliberadamente junto con Prometeo en una gran explosión termonuclear. Ya no era necesario prohibir la tecnología moderna, ya que todo había desaparecido. Como resultado, la humanidad se procuró mil años más sin tener que preocuparse por la IA o el totalitarismo.

Aunque en menor medida, ya se ha dado alguna vez una vuelta atrás: por ejemplo, algunas de las tecnologías cuyo uso estaba ampliamente extendido durante el Imperio romano fueron en gran medida olvidadas durante casi un milenio hasta que reaparecieron durante el Renacimiento. La trilogía de la *Fundación* de Isaac Asimov gira en torno al «plan Seldon» para reducir la duración de un periodo de vuelta atrás de 30.000 a 1.000 años. Con una ingeniosa planificación, sería posible hacer lo contrario y alargar en lugar de acortar un periodo de vuelta atrás, por ejemplo suprimiendo todo conocimiento sobre agricultura. No obstante, para desgracia de los entusiastas de la vuelta atrás, es poco probable que este escenario pudiese prolongarse indefinidamente sin que la humanidad desarrollara tecnologías avanzadas o bien se extinguiese. Pecaríamos de ingenuos si diésemos por seguro que los seres humanos de dentro de cien millones de años se parecerán a los actuales, puesto que de momento no hemos existido como especie más que un 1 % de ese tiempo. Además, una humanidad sin tecnología avanzada sería un blanco fácil e indefenso, esperando a ser exterminado por el próximo impacto de un asteroide capaz de abrasar el planeta u otra megacalamidad provocada por la madre naturaleza. Ciertamente, no duraremos mil millones de años, pues transcurrido ese tiempo el aumento gradual de la temperatura del Sol habrá

calentado la Tierra lo suficiente para que toda el agua en estado líquido se haya evaporado.

AUTODESTRUCCIÓN

Tras considerar los problemas que podría causar la tecnología futura, conviene considerar también los que la falta de esa tecnología podría provocar. En esa línea, exploremos escenarios donde la superinteligencia no llega a crearse porque la humanidad se extingue por otros medios.

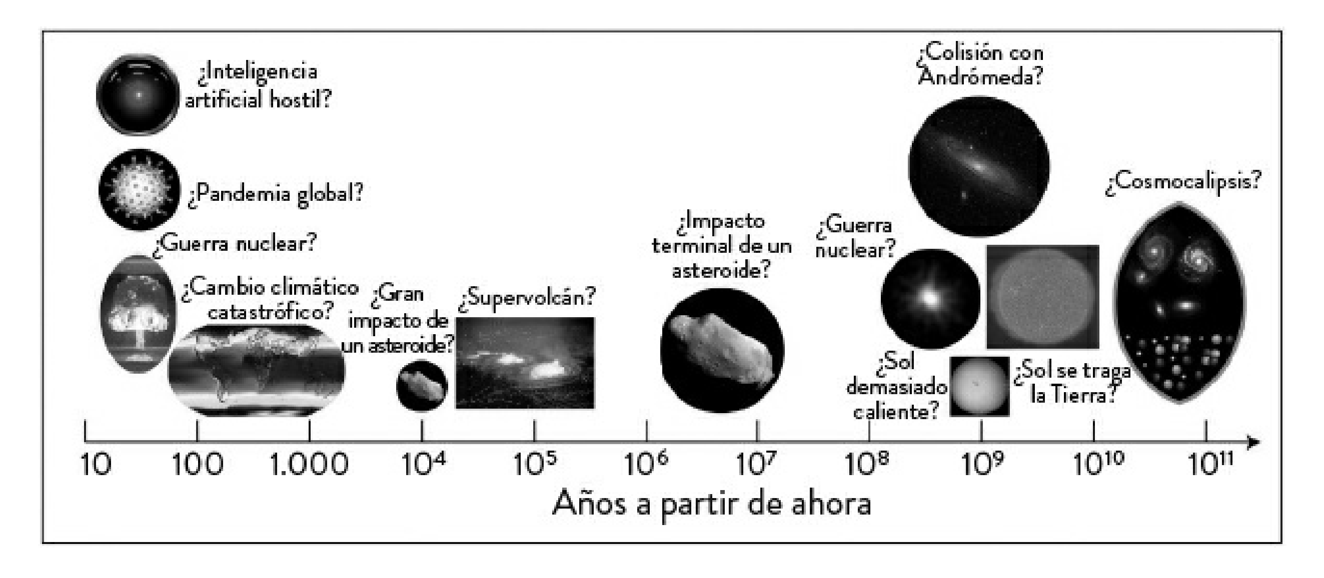


FIGURA 5.1. Ejemplos de lo que podría destruir la vida tal como la conocemos o limitar permanentemente su potencial. Aunque es probable que el universo en sí perdure al menos decenas de miles de millones de años, el Sol abrasará la Tierra en unos mil millones de años y luego se la tragará, a menos que la desplacemos a una distancia segura, y nuestra galaxia colisionará con su vecina dentro de unos 3.500 millones de años. Aunque no sabemos exactamente cuándo, podemos predecir con certeza que mucho antes de esto algún asteroide impactará contra la Tierra y los supervolcanes provocarán inviernos durante los que no se verá el sol en todo el año. Podemos utilizar la tecnología para resolver todos estos problemas o para crear otros nuevos, como el cambio climático, la guerra nuclear, las pandemias de diseño o una IA malograda.

¿Cómo podríamos lograr eso? La estrategia más simple es «sentarse a esperar». Aunque en el capítulo siguiente veremos cómo podemos resolver problemas como los impactos de asteroides y la ebullición de los océanos, todas estas soluciones requieren tecnología que aún no hemos desarrollado,

por lo que, a menos que nuestra tecnología avance mucho más allá de su nivel actual, la madre naturaleza nos extinguirá mucho antes de que hayan transcurrido otros mil millones de años. Como dijo el famoso economista John Maynard Keynes: «A la larga, todos estaremos muertos».

Desgraciadamente, también hay formas en que podríamos autodestruirnos mucho antes como consecuencia de nuestra estupidez colectiva. ¿Por qué iba a cometer nuestra especie un suicidio colectivo, también denominado omnicidio, si casi nadie lo desea? Con nuestro nivel actual de inteligencia y madurez emocional, los humanos tenemos especial habilidad para los errores de cálculo, los malentendidos y la incompetencia, y, en consecuencia, nuestra historia está llena de accidentes, guerras y otras calamidades que, en retrospectiva, prácticamente nadie deseaba. Los economistas y los matemáticos han desarrollado explicaciones elegantes basadas en la teoría de juegos para explicar cómo se puede incentivar a las personas a realizar acciones que, en última instancia, tienen un resultado catastrófico para todos.[74]

Guerra nuclear: caso práctico de imprudencia humana

Podríamos pensar que cuanto más haya en juego más cuidadosos seríamos, pero un análisis más detallado del mayor riesgo que nuestra tecnología actual hace posible, el de una guerra termonuclear global, no resulta tranquilizador. Hemos tenido que depender de la suerte para sortear una lista vergonzosamente larga de incidentes causados por todo tipo de cosas: averías informáticas, cortes de luz, información defectuosa, errores de navegación, accidentes aéreos de bombarderos, explosiones de satélites, etcétera.[75] De hecho, de no ser por las actuaciones heroicas de determinados individuos —como Vasili Arjípov y Stanislav Petrov— puede que ya hubiésemos tenido una guerra nuclear global. Dado nuestro historial, considero realmente increíble que la probabilidad anual de guerra nuclear accidental sea tan baja como del uno por mil si perseveramos en nuestro comportamiento actual, en cuyo caso la probabilidad de que esa guerra tenga lugar en los próximos 10.000 años es de más del $1 - 0{,}999^{10000} \approx 99{,}995$ %.

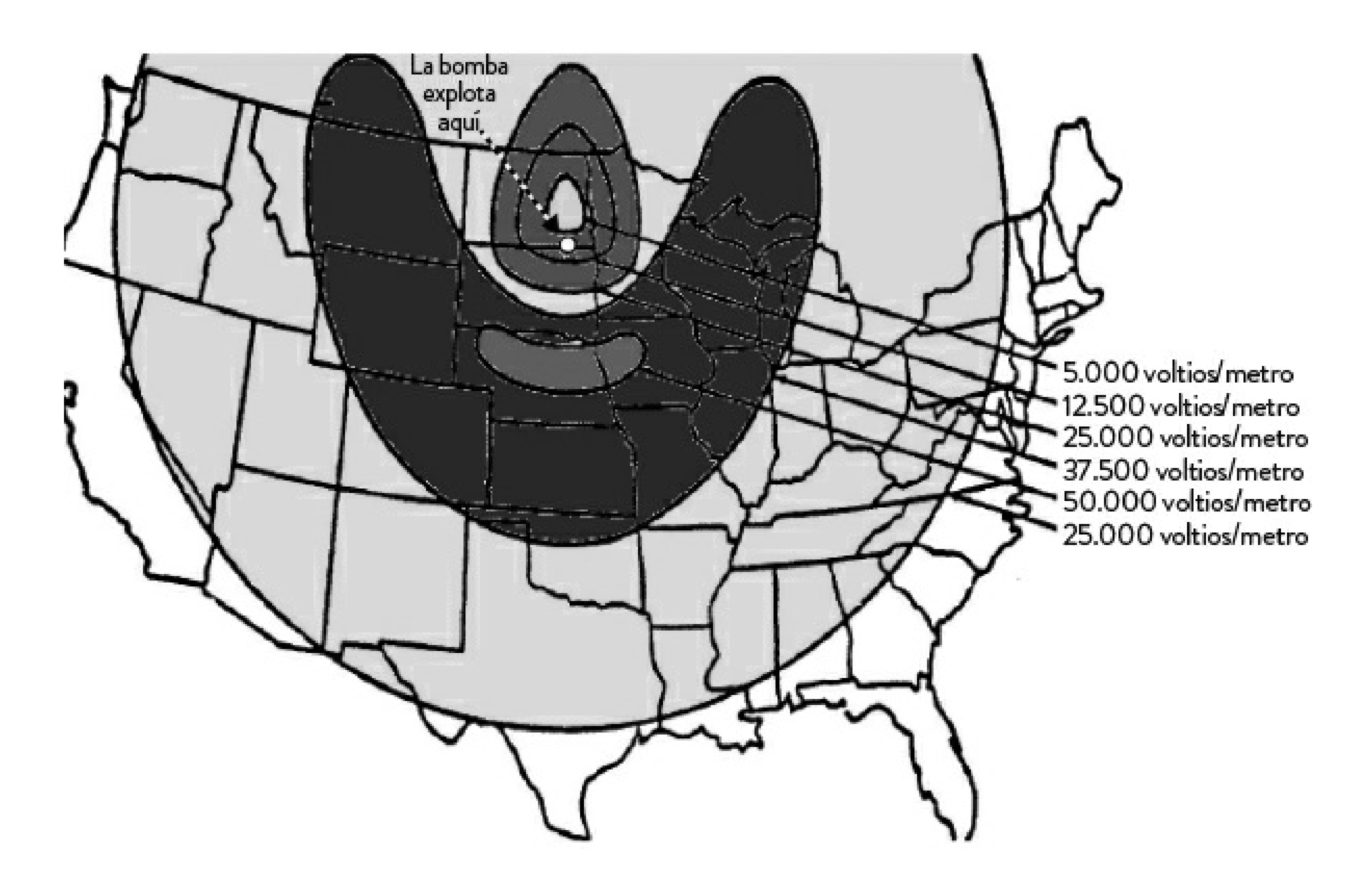


FIGURA 5.2. La explosión de una sola bomba de hidrógeno a 400 kilómetros por encima de la superficie puede generar un potente pulso electromagnético capaz de paralizar las tecnologías que necesiten electricidad en una gran extensión. Al desplazar el punto de detonación al sureste, la zona en forma de plátano que supera los 37.500 voltios por metro podría cubrir la mayor parte de la costa Este. Reproducido a partir del informe AD-A278230 (sin clasificar) del ejército estadounidense.

Para apreciar en toda su magnitud la imprudencia humana, debemos tomar conciencia de que nos lanzamos a la apuesta nuclear antes incluso de evaluar cuidadosamente los riesgos. En primer lugar, se habían subestimado los riesgos de radiación, y, solo en Estados Unidos, se han pagado compensaciones por valor de más de 2.000 millones de dólares a las víctimas de la exposición a radiación por la manipulación de uranio y por las pruebas nucleares.[76]

En segundo lugar, posteriormente se descubrió que la detonación deliberada de una bomba de hidrógeno a cientos de kilómetros sobre la superficie terrestre generaría un potente pulso electromagnético (EMP, por sus siglas en inglés) capaz de inutilizar la red eléctrica y los dispositivos electrónicos en grandes extensiones (figura 5.2), y así dejar la infraestructura paralizada, las carreteras bloqueadas con vehículos inutilizados y condiciones

poco propicias para la supervivencia. Por ejemplo, un informe de la comisión estadounidense sobre EMP alertó de que «la infraestructura hídrica es una gigantesca máquina, movida en parte gracias a la gravedad, pero principalmente mediante electricidad», y que la falta de agua podría provocar la muerte en tres o cuatro días.[77]

En tercer lugar, no se pensó en la posibilidad de que se produjese un invierno nuclear hasta transcurridas cuatro décadas, cuando ya se habían desplegado 63.000 bombas de hidrógeno (¡uy!). Con independencia de qué ciudades fuesen las que ardiesen en llamas, podrían dispersarse cantidades ingentes de humo por todo el mundo al alcanzar la troposfera superior, y bloquear una proporción suficiente de la luz solar para transformar los veranos en inviernos, como cuando, en el pasado, un asteroide o supervolcán provocaron una extinción masiva. Cuando científicos tanto estadounidenses como soviéticos hicieron sonar la voz de alarma en los años ochenta, contribuyeron a la decisión de Ronald Reagan y Mijaíl Gorbachov de comenzar a reducir sus arsenales.[78] Por desgracia, cálculos más precisos dibujan un panorama aún más sombrío: la figura 5.3 muestra un enfriamiento de alrededor de 20 grados centígrados en buena parte de las regiones agrícolas centrales de Estados Unidos, Europa, Rusia y China (y de 35º C en algunas zonas de Rusia) durante los dos primeros veranos, y de aproximadamente la mitad de esa temperatura incluso transcurrida toda una década.(17) ¿Qué significa esto en un lenguaje comprensible? No hay que tener mucha experiencia agrícola para saber que temperaturas veraniegas cercanas a cero durante años acabarían con casi toda nuestra producción de alimentos. Es difícil predecir con precisión qué sucedería una vez que miles de las mayores ciudades del planeta fuesen reducidas a escombros y la infraestructura global se viniese abajo, pero la pequeña proporción de todos los humanos que no sucumbiese al hambre, la hipotermia o las enfermedades tendría que hacer frente a hordas itinerantes armadas en busca desesperada de comida.

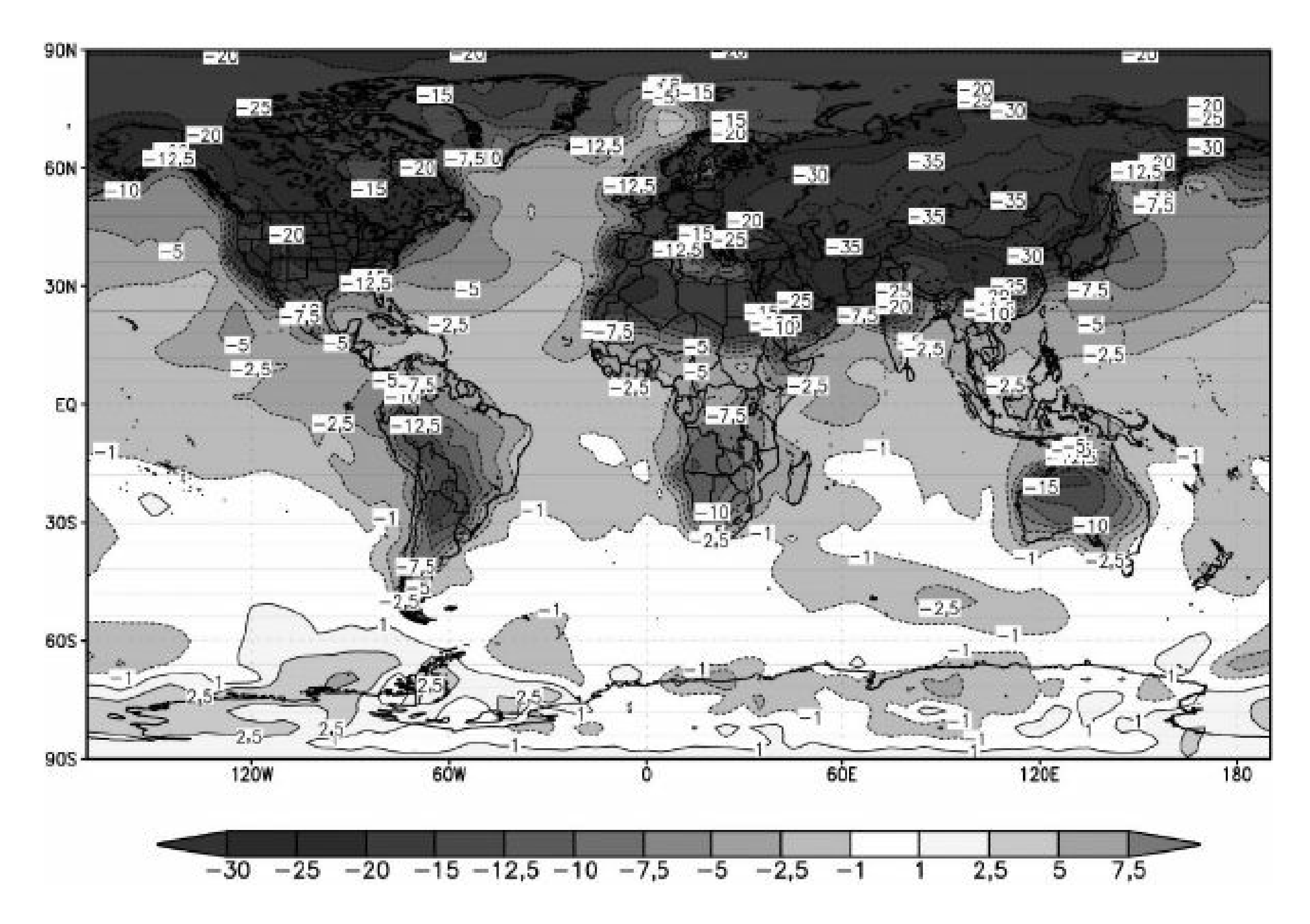


FIGURA 5.3. Enfriamiento medio (en grados centígrados) durante los dos primeros veranos tras una guerra nuclear total entre Estados Unidos y Rusia. Reproducido con permiso de Alan Robock.[79]

He entrado en este grado de detalle sobre la guerra nuclear global para dejar muy clara la idea fundamental de que ningún líder mundial razonable la desearía, a pesar de lo cual podría no obstante suceder accidentalmente. Esto significa que no podemos confiar en que nuestros congéneres nunca cometerán un omnicidio: el hecho de que nadie lo desee no basta para evitarlo.

Dispositivos del apocalipsis

¿Deberíamos los humanos cometer un omnicidio? Incluso aunque una guerra nuclear global podría exterminar al 90 % de los humanos, la mayoría de los científicos estiman que no mataría al 10 % restante, y por lo tanto no provocaría nuestra extinción. Por otra parte, los antecedentes de la radiación

nuclear, el EMP nuclear y el invierno nuclear demuestran que los peligros más grandes pueden ser aquellos que aún no hemos imaginado. Es extraordinariamente difícil prever todos los aspectos de la situación resultante y cómo el invierno nuclear, la destrucción de las infraestructuras, los elevados niveles de mutación genética y las hordas armadas desesperadas podrían interactuar con otros problemas como nuevas pandemias, el colapso del ecosistema y otros efectos que aún ni siquiera imaginamos. Por lo tanto, mi evaluación personal es que, aunque no sea grande la probabilidad de que mañana se produzca una guerra nuclear que provoque la extinción humana, tampoco podemos concluir con seguridad que sea nula.

La probabilidad de que se produzca un omnicidio aumenta si modernizamos las armas nucleares actuales hasta crear deliberadamente un dispositivo del Apocalipsis. Propuesto en 1960 por el estratega de RAND Herman Kahn, y popularizado en la película de Stanley Kubrick *¿Teléfono rojo? Volamos hacia Moscú*, un dispositivo del Apocalipsis lleva el paradigma de la destrucción asegurada de ambos países a sus últimas consecuencias. Es el elemento de disuasión perfecto: una máquina que automáticamente toma represalias contra cualquier ataque enemigo matando a toda la humanidad.

Un candidato a dispositivo del Apocalipsis es un gran arsenal subterráneo de las llamadas bombas nucleares saladas, preferiblemente enormes bombas de hidrógeno rodeadas de cantidades ingentes de cobalto. El físico Leo Szilard ya argumentó en 1950 que esto podría acabar con todos los habitantes de la Tierra: las explosiones de la bomba de hidrógeno convertirían el cobalto en radiactivo y lo lanzarían hasta la estratosfera, y su vida media de cinco años es suficiente para extenderse por toda la Tierra (sobre todo si se colocasen dos dispositivos del Apocalipsis gemelos en hemisferios opuestos), pero lo suficientemente corta para generar radiación de una intensidad letal. Las informaciones de la prensa sugieren que, hoy en día, por primera vez se están fabricando bombas de cobalto. Las oportunidades de un omnicidio podrían incrementarse añadiendo bombas diseñadas para provocar un invierno nuclear, al maximizar los aerosoles de larga vida en la estratosfera. Una importante ventaja de un dispositivo del Apocalipsis es que es mucho más barato que un elemento de disuasión nuclear convencional: dado que no es necesario lanzar bombas, no nos vemos obligados a construir costosos sistemas de misiles, y las propias bombas son más baratas de construir ya que

no tienen que ser lo suficientemente ligeras y compactas como para caber en los misiles.

Otra posibilidad es el futuro descubrimiento de un dispositivo biológico del Apocalipsis: una bacteria o virus diseñado para matar a todos los humanos. Si su transmisibilidad fuera lo suficientemente alta y su periodo de incubación lo bastante largo, la mayoría de la gente podría resultar infectada antes de enterarse de su existencia y de tomar contramedidas. Hay un argumento militar para la fabricación de tal arma biológica incluso aunque no sirviera para matar a todo el mundo: el dispositivo del día del Apocalipsis más efectivo es aquel que combina armas nucleares, biológicas y de otro tipo para maximizar las posibilidades de disuadir al enemigo.

Armas con IA

Una tercera vía tecnológica hacia el omnicidio puede involucrar armas relativamente tontas con IA. Supongamos que una superpotencia construye miles de millones de esos drones de ataque del tamaño de abejorros que vimos en el capítulo 3 y los utiliza para matar a todo el mundo salvo a sus propios ciudadanos y a sus aliados, identificados de forma remota por una etiqueta de identificación de radiofrecuencia (RFID, por sus siglas en inglés), como lo están la mayoría de los productos de supermercado de hoy. Estas etiquetas podrían distribuirse a todos los ciudadanos para usarlas en pulseras o como implantes transcutáneos, como comentamos en la sección sobre el escenario totalitario. Esto sin duda impulsaría a una superpotencia opuesta a construir algo análogo. Si una guerra estallase accidentalmente, morirían todos los humanos, incluso los miembros de tribus remotas no alineadas con una u otra superpotencia, porque nadie llevaría ambos tipos de identificación. Si se combinase lo anterior con un dispositivo del Apocalipsis nuclear y biológico, aumentarían aún más las posibilidades de que el omnicidio tuviera éxito.

¿QUÉ QUIERE USTED?

Comenzamos este capítulo reflexionando sobre el lugar al que deseamos que

nos conduzca la actual carrera hacia la IAG. Ahora que hemos explorado juntos una amplia gama de escenarios, ¿cuáles le parecen interesantes y cuáles cree que deberíamos tratar de evitar? ¿Tiene algún claro favorito? Por favor, comparta sus ideas conmigo y con los demás lectores en <http://AgeOfAi.org> y participe en la conversación.

Evidentemente, los escenarios que hemos cubierto no deberían entenderse como una lista completa; además, muchos de ellos son apenas esbozos. Pero sí he intentado ser incluyente y abarcar todo el espectro, desde escenarios con tecnología avanzada hasta otros con una menos desarrollada, o incluso sin tecnología. También he procurado describir todos los temores y esperanzas más habituales expresados en la literatura.

Una de las partes más entretenidas al escribir este libro ha sido escuchar lo que mis amigos y colegas opinan sobre estos escenarios, y me ha divertido comprobar que no hay consenso alguno. La única cosa con la que todo el mundo está de acuerdo es que las opciones son más matizadas de lo que inicialmente podría parecer: las personas a las que les gusta alguno de los escenarios tienden a encontrar al mismo tiempo criticable algún aspecto o aspectos del mismo. Para mí, esto significa que los seres humanos necesitamos continuar y profundizar esta conversación sobre nuestros objetivos para el futuro, para saber hacia dónde queremos ir. Las posibilidades futuras para la vida en el cosmos son impresionantemente grandiosas, no las desperdiciemos dejándonos ir a la deriva como un barco sin timonel, sin tener ni idea de hacia dónde queremos dirigirnos.

¿Cómo son de grandiosas estas posibilidades futuras? Por avanzada que sea nuestra tecnología, la capacidad de la vida 3.0 para mejorar y extenderse a través del cosmos estará limitada por las leyes de la física. ¿Cuáles son estos límites últimos, durante los próximos miles de millones de años? ¿Está el universo lleno de vida extraterrestre, o estamos solos aquí? ¿Qué sucedería si se encontrasen dos civilizaciones cósmicas en expansión? Abordaremos estas fascinantes preguntas en el próximo capítulo.

CONCLUSIONES

- La carrera actual hacia IAG puede desembocar en una variedad fascinantemente amplia de escenarios posteriores a su desarrollo para los próximos milenios.

- La superinteligencia puede coexistir pacíficamente con los humanos, ya sea porque se ve forzada a ello (escenario de dios esclavizado) o porque es una «IA amigable» que quiere hacerlo (escenarios de utopía libertaria, dios protector, dictador benévolo y cuidador de zoo).
- El desarrollo de la superinteligencia se puede evitar mediante una IA (escenario del guardián) o por los humanos (escenario de 1984), olvidando deliberadamente la tecnología (escenario de vuelta atrás) o por falta de incentivos para construirla (por ejemplo, escenario de utopía igualitaria).
- La humanidad puede extinguirse y ser reemplazada por la IA (escenarios de dominadores y descendientes) o por ninguna otra cosa (escenario de autodestrucción).
- No existe absolutamente ningún consenso en torno a cuáles de estos escenarios son deseables, si es que alguno lo es, y todos ellos tienen sus inconvenientes. Esto hace que sea aún más importante continuar y profundizar la conversación en torno a nuestros objetivos futuros, para que no nos desviemos sin ser conscientes de ello ni nos aventuremos en una dirección equivocada.

6
NUESTRA HERENCIA CÓSMICA
LOS PRÓXIMOS MIL MILLONES DE AÑOS Y MÁS ALLÁ

> *Nuestra especulación desemboca en una supercivilización, la síntesis de toda la vida en el sistema solar, que se mejora y se extiende constantemente, difundiéndose a distancias cada vez mayores del Sol, transformando la materia inerte en mente.*
>
> HANS MORAVEC, *El hombre mecánico*

En mi opinión, el descubrimiento científico más estimulante de la historia es que hemos subestimado espectacularmente el potencial futuro de la vida. Nuestros sueños y aspiraciones no tienen por qué limitarse a vivir cien años lastrados por enfermedades, pobreza y confusión, sino que, con la ayuda de la tecnología, la vida tiene el potencial de florecer durante miles de millones de años, no solo aquí, en nuestro sistema solar, sino a lo largo y ancho de un universo mucho más inmenso y fascinante de lo que nuestros antepasados imaginaron. Ni siquiera el cielo es el límite.

Esta es una noticia ilusionante para una especie que ha sentido la necesidad de ampliar sus horizontes a lo largo de los siglos. Los Juegos Olímpicos celebran la ampliación de los límites de la fuerza, la velocidad, la agilidad y la resistencia. La ciencia celebra la ampliación de los límites del conocimiento y la comprensión. La literatura y el arte celebran la ampliación de los límites de la creación de experiencias hermosas o enriquecedoras. Muchas personas, organizaciones y países celebran el incremento de los recursos, el territorio y la longevidad. Dada nuestra obsesión con la superación de los límites, parece apropiado que el libro sujeto a *copyright* más vendido de todos los tiempos sea *El libro Guinness de los récords*.

Si la tecnología puede hacer añicos las barreras a las que tradicionalmente pensábamos que estaba sometida la vida, ¿cuáles son los límites últimos? ¿Qué proporción del cosmos puede cobrar vida? ¿Hasta dónde y hasta cuándo puede llegar la vida? ¿De cuánta materia puede la vida hacer uso, y cuánta energía, información y computación puede extraer de ella? Estos límites

últimos no vienen definidos por nuestro grado de comprensión del universo, sino por las leyes físicas. Lo cual, paradójicamente, hace que en ciertos aspectos resulte más fácil analizar el futuro de la vida a largo que a corto plazo.

Si nuestra historia cósmica de 13.800 millones de años se comprimiese en una semana, todo el drama de 10.000 años de duración de los dos últimos capítulos habría acabado en menos de medio segundo. Esto significa que, aunque no podemos predecir si la explosión de inteligencia tendrá lugar y cómo se desarrollará, ni cuáles serán sus consecuencias inmediatas, toda esta agitación no es más que un breve destello en la historia cósmica, cuyos detalles no afectan a los límites últimos de la vida. Si la vida tras la explosión está tan obsesionada con superar las barreras como lo estamos ahora los humanos, desarrollará tecnología para sobrepasar realmente dichos límites, porque podrá hacerlo. En este capítulo, veremos cuáles son, y así nos haremos una idea de cómo será el futuro de la vida a largo plazo. Dado que estos límites se basan en nuestra comprensión actual de la física, deben considerarse como una cota inferior a lo que es posible: los descubrimientos científicos futuros pueden abrir oportunidades para mejorar aún más.

Pero ¿realmente sabemos si la vida futura será tan ambiciosa? No lo sabemos: tal vez se vuelva tan indolente como un adicto a la heroína o a la televisión, y se dedique a ver un programa del corazón tras otro. Sin embargo, hay motivos para sospechar que la ambición es un rasgo bastante genérico de la vida avanzada. Casi con independencia de lo que intente maximizar, ya sea la inteligencia, la longevidad, el conocimiento o las experiencias interesantes, le harán falta recursos. Por lo tanto, tendrá incentivos para desarrollar su tecnología hasta los límites últimos, para sacar el máximo partido a los recursos de que disponga. Tras esto, la única manera de seguir mejorando será obtener más recursos, expandiéndose a zonas cada vez más amplias del cosmos.

Además, la vida puede surgir de manera independiente en varios lugares del universo. De ser así, las civilizaciones poco ambiciosas serían cósmicamente irrelevantes, mientras que las formas de vida más ambiciosas se irían haciendo por una proporción cada vez mayor de la herencia cósmica. Así pues, la selección natural tendrá lugar en una escala cósmica y, transcurrido un tiempo, casi toda la vida que existirá será vida ambiciosa. En resumen, si nos interesa saber en qué medida puede cobrar vida el universo

en última instancia, debemos estudiar los límites que las leyes de la física imponen a la ambición. Hagámoslo. Exploraremos primero los límites de lo que se puede hacer con los recursos (materia, energía, etcétera) que tenemos en el sistema solar, y a continuación veremos cómo obtener más recursos a través de la exploración y la colonización cósmicas.

APROVECHAR AL MÁXIMO NUESTROS RECURSOS

Mientras que los supermercados actuales y los mercados de materias primas venden decenas de miles de productos que podríamos llamar «recursos», la vida futura, una vez haya alcanzado el límite tecnológico, necesitará un solo recurso fundamental: la llamada materia bariónica, es decir, cualquier cosa compuesta por átomos o por sus componentes (quarks y electrones). Independientemente de la forma en que se encuentre esta materia, la tecnología avanzada podrá reorganizarla para obtener cualquier sustancia u objeto que desee, incluidos centrales eléctricas, ordenadores y formas de vida avanzadas. Comencemos por examinar los límites existentes sobre la energía de la que se alimenta la vida avanzada y sobre el procesamiento de la información que le permite pensar.

Construir esferas de Dyson

Si hablamos del futuro de la vida, uno de los visionarios más optimistas es Freeman Dyson. He tenido el honor y el placer de conocerlo durante las últimas dos décadas, pero la primera vez que lo vi me puse nervioso. Yo era un joven postdoc que estaba comiendo con mis amigos en el comedor del Instituto de Estudios Avanzados en Princeton, y de repente este físico de fama mundial, que había tratado a Einstein y a Gödel, se acercó, se presentó, y nos preguntó si podía sentarse con nosotros. Enseguida consiguió que me relajase, al explicarme que prefería almorzar con gente joven que con profesores viejos y aburridos. A pesar de que tiene noventa y tres años cuando escribo estas palabras, Freeman sigue siendo aún más joven de espíritu que la mayoría de las personas que conozco, y el pícaro destello juvenil en sus ojos revela que no le importan las formalidades, las jerarquías

académicas o la sabiduría convencional. Cuanto más audaz es la idea, más le entusiasma.

Cuando hablé con él sobre el uso de energía, se burló de la falta de ambición de los humanos, señalando que podríamos satisfacer todas las necesidades energéticas mundiales actuales aprovechando la luz solar que incide sobre una superficie de menos del 0,5 % de la extensión del desierto del Sáhara. Pero ¿por qué limitarse a eso? ¿Por qué limitarse incluso a capturar toda la luz del Sol que incide sobre la Tierra, y dejar que la mayor parte de ella se desperdicie al ser emitida hacia el espacio vacío? ¿Por qué no simplemente aprovechar toda la producción de energía solar para la vida?

Inspirado en el clásico de la ciencia ficción *Hacedor de estrellas*, publicado por Olaf Stapledon en 1937, y sus anillos de mundos artificiales orbitando alrededor de sus respectivas estrellas, Freeman Dyson publicó en 1960 una descripción de lo que acabaría por conocerse como una *esfera de Dyson*.[80] La idea de Freeman era reconvertir Júpiter en una biosfera en forma de cáscara esférica alrededor del Sol, donde nuestros descendientes podrían prosperar, al tener acceso a cien mil millones de veces más biomasa y un billón de veces más energía que la que la humanidad usa hoy en día.[81] Dyson argumentó que era el siguiente paso natural: «Es de esperar que, al cabo de unos pocos miles de años de entrar en la fase de desarrollo industrial, cualquier especie inteligente estuviera ocupando una biosfera artificial que rodease por completo la estrella alrededor de la que orbita». Si viviésemos en el interior de una esfera de Dyson, no habría noches: siempre veríamos el Sol directamente sobre nuestras cabezas, y, por todo el cielo, su luz reflejándose en el resto de la biosfera, como ahora la vemos reflejada en la Luna durante el día. Si quisiésemos ver las estrellas, tendríamos que «subir» y asomarnos al universo desde el exterior de la esfera de Dyson.

Una manera que no requiere tecnología tan avanzada de construir una esfera de Dyson parcial consiste en colocar un anillo de hábitats en órbita circular alrededor del Sol. Para rodear por completo el Sol, se podrían añadir anillos que orbitasen según diferentes ejes y distancias ligeramente distintas, para evitar colisiones. Para impedir que estos anillos, que se moverían a gran velocidad, no pudiesen conectarse entre sí, lo que dificultaría el transporte y la comunicación, en lugar de lo anterior se podría construir una esfera de Dyson monolítica y estacionaria, en la que la atracción gravitatoria hacia el Sol se compensase con la presión hacia fuera que ejerce la radiación

procedente de la estrella (una idea propuesta por primera vez por Robert L. Forward y por Colin McInnes). La esfera podría construirse gradualmente añadiendo «estátites», satélites estacionarios que contrarrestan la gravedad solar con la presión de la radiación, en lugar de con fuerzas centrífugas. Estas dos fuerzas decaen de forma proporcional al cuadrado de la distancia al Sol, lo que significa que, si pueden equilibrarse a una distancia del Sol, también podrán estar equilibradas a cualquier otra distancia, lo que da libertad para estacionar en cualquier lugar del sistema solar. Los estátites deberían estar formados por láminas muy ligeras, con una masa de solo 0,77 gramos por metro cuadrado, lo cual es aproximadamente cien veces más ligero que el papel, aunque esto no debería ser un obstáculo insalvable. Por ejemplo, una lámina de grafeno (una sola capa de átomos de carbono en un patrón hexagonal que se asemeja a un alambre de gallinero) pesa mil veces menos que ese límite. Si la esfera de Dyson está diseñada para reflejar en lugar de absorber la mayor parte de la luz solar, la intensidad total de la luz que rebota en su interior aumentará de forma drástica, incrementando aún más la presión de radiación y la cantidad de masa que la esfera puede soportar. Muchas otras estrellas tienen una luminosidad mil e incluso un millón de veces mayor que el Sol, y por lo tanto pueden soportar esferas de Dyson estacionarias proporcionalmente más pesadas.

Si se desea construir una esfera de Dyson rígida mucho más pesada aquí en nuestro sistema solar, entonces para que soporte la gravedad del Sol se necesitarán materiales ultrafuertes capaces de resistir presiones decenas de miles de veces mayores que las que soportan los cimientos de los rascacielos más altos sin licuarse o arrugarse. Para ser duradera, la esfera Dyson tendría que ser dinámica e inteligente, ajustar constantemente su posición y su forma en respuesta a las perturbaciones, y abrir ocasionalmente grandes agujeros para permitir el paso de molestos asteroides y cometas. Otra posibilidad sería usar un sistema de detección y desviación para gestionar esos objetos ajenos al sistema, que podría incluso descomponerlos para usar los materiales de los que estuviesen hechos.

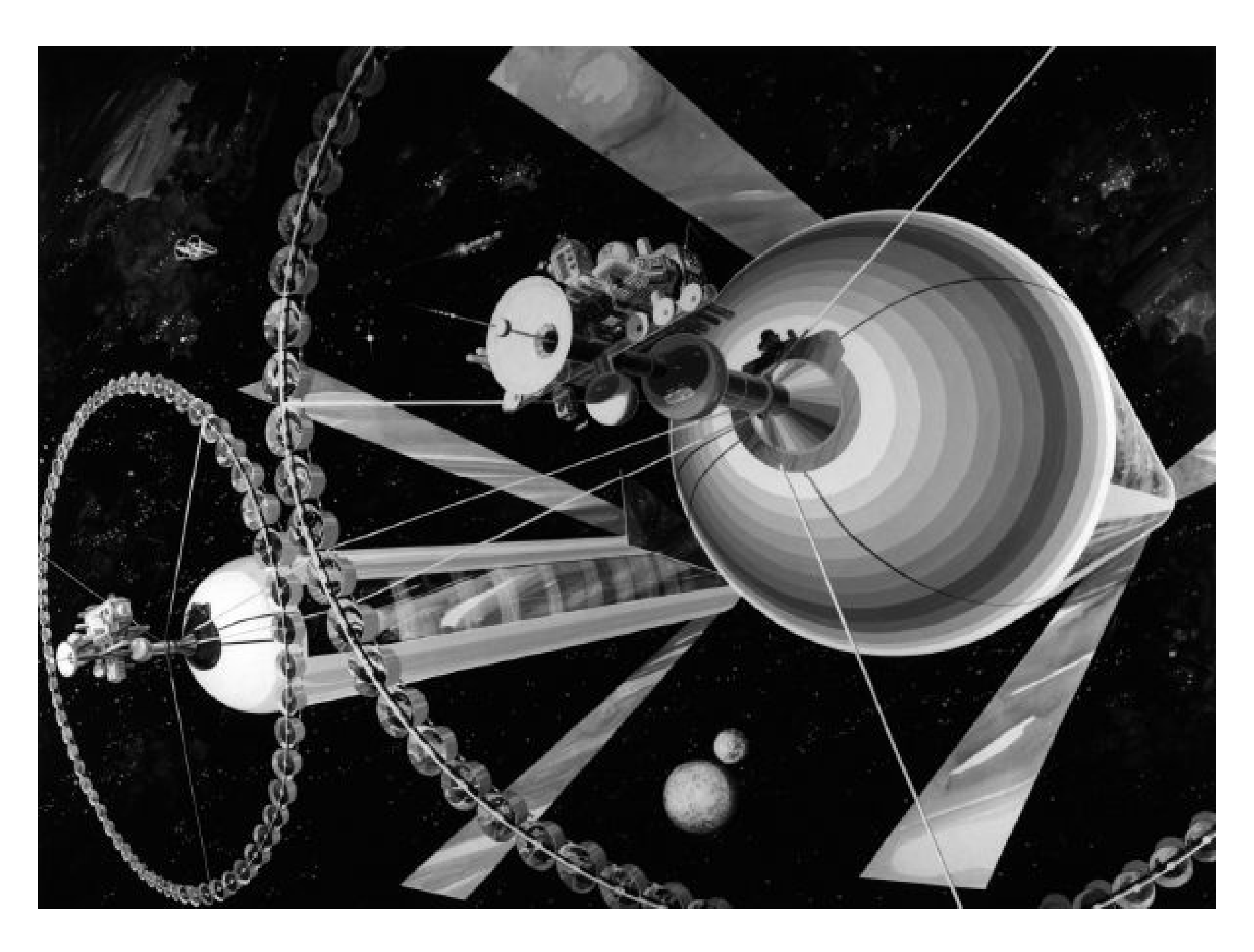


FIGURA 6.1. Un par de cilindros de O'Neill contrarrotatorios pueden proporcionar cómodos hábitats humanos parecidos a la Tierra si orbitan alrededor del Sol de tal forma que siempre apunten directamente hacia él. La fuerza centrífuga de su rotación proporciona gravedad artificial, y tres espejos plegables irradian luz solar en el interior con un periodo de veinticuatro horas, sumando el día y la noche. Los hábitats más pequeños dispuestos en un anillo están destinados a la agricultura. Imagen cortesía de Rick Guidice/NASA.

Para los humanos de hoy en día, la vida en el interior o sobre la superficie de una esfera de Dyson sería, en el mejor de los casos, desconcertante, y, en el peor, imposible, pero eso no tiene por qué ser obstáculo para que las futuras formas de vida, biológicas o no, prosperen allí. La variante de la esfera en órbita no tendría básicamente ninguna gravedad, y si uno caminase sobre una esfera de tipo estacionario, solo podría hacerlo por el exterior (la superficie más alejada del Sol) sin caerse, con una gravedad unas diez mil veces más débil que la habitual. No habría campo magnético (a menos que se hubiese construido uno) que protegiese de las partículas peligrosas procedentes del Sol. Lo bueno es que en una esfera de Dyson del tamaño de

la órbita actual de la Tierra tendríamos una superficie habitable aproximadamente quinientos millones de veces mayor que la actual.



FIGURA 6.2. Vista interior de uno de los cilindros O’Neill de la figura anterior. Si su diámetro es de 6,4 kilómetros y da una vuelta completa cada dos minutos, las personas situadas en la superficie experimentarán la misma gravedad aparente que en la Tierra. El Sol está a nuestras espaldas, pero aparece arriba debido a un espejo colocado fuera del cilindro que se abate por la noche. Las ventanas herméticas evitan que la atmósfera escape del cilindro. Imagen cortesía de Rick Guidice/NASA.

Si se desean hábitats humanos más similares a la Tierra, la buena noticia es que son mucho más fáciles de construir que una esfera de Dyson. Por ejemplo, las figuras 6.1 y 6.2 muestran un diseño de hábitat cilíndrico propuesto por el físico estadounidense Gerard K. O’Neill, que incluye gravedad artificial, blindaje contra los rayos cósmicos, un ciclo diurno-nocturno de veinticuatro horas y una atmósfera y ecosistemas similares a los terrestre. Dichos hábitats podrían orbitar libremente dentro de una esfera de

Dyson, o bien, retocando su forma, podrían sujetarse al exterior de la misma.

Construir mejores centrales eléctricas

A pesar de que las esferas de Dyson son eficientes desde el punto de vista energético según los estándares de ingeniería actuales, no se acercan ni remotamente a los límites que marcan las leyes físicas. Einstein nos enseñó que, si fuésemos capaces de convertir la masa en energía con un 100 % de eficiencia,(18) entonces una cantidad de masa m nos proporcionaría una cantidad de energía E dada por su famosa fórmula $E = mc^2$, donde c es la velocidad de la luz. Esto significa que, dado que c es enorme, una pequeña cantidad de masa puede producir una gran cantidad de energía. Si tuviéramos un suministro abundante de antimateria (cosa que no sucede), sería fácil construir una central eléctrica con una eficiencia del 100 %: bastaría con verter una cucharadita de anti-agua en agua normal para liberar la energía equivalente a 200.000 toneladas de TNT, la cantidad de energía que libera una bomba de hidrógeno típica, suficiente para cubrir las necesidades energéticas de todo el planeta durante unos siete minutos.

Por su parte, las formas más comunes de generar energía en la actualidad son tremendamente ineficientes, como se resume en la tabla 6.1 y en la figura 6.3. La digestión de un caramelo tiene una eficiencia de tan solo el 0,00000001 %, en el sentido de que libera apenas una diezbillonésima parte de la energía mc^2 que contiene. Si nuestro estómago tuviese siquiera una eficiencia del 0,001 %, entonces nos bastaría con comer una sola comida en lo que nos queda de vida. En comparación con la comida, la combustión de carbón y de gasolina es solo tres y cinco veces más eficiente, respectivamente. Los reactores nucleares actuales son mucho mejores a la hora de dividir átomos de uranio por medio de la fisión, pero aun así no pueden extraer más que el 0,08 % de su energía. El reactor nuclear en el núcleo del Sol es un orden de magnitud más eficiente que los construidos por el ser humano, y extrae el 0,7 % de la energía del proceso de fusión del hidrógeno para producir helio. Sin embargo, incluso si encerramos al Sol en una esfera de Dyson perfecta, nunca convertiremos más del 0,08 % de la masa del Sol en energía útil, porque, una vez que este haya consumido

aproximadamente una décima parte de su combustible de hidrógeno, terminará su vida como una estrella normal, se expandirá a una estrella gigante roja y comenzará a morir. Las cosas tampoco son mucho mejores en el caso de otras estrellas: la proporción de su hidrógeno consumida durante su vida útil oscila entre el 4 % para estrellas muy pequeñas y alrededor del 12 % para las más grandes. Si perfeccionásemos un reactor de fusión artificial que nos permitiese fusionar el 100 % de todo el hidrógeno a nuestra disposición, todavía seríamos incapaces de superar la eficiencia ridículamente baja del 0,7 % del proceso de fusión. ¿Cómo podemos hacerlo mejor?

MÉTODO	EFICIENCIA
Digerir un caramelo	0,00000001 %
Quemar carbón	0,00000003 %
Quemar gasolina	0,00000005 %
Fisionar uranio-235	0,08 %
Usar una esfera de Dyson hasta que muera el Sol	0,08 %
Fusión de hidrógeno en helio	0,7 %
Motor de agujero negro rotatorio	29 %
Esfera de Dyson alrededor de un cuásar	42 %
Esfalerizador	50 % (?)
Evaporación de un agujero negro	90 %*

TABLA 6.1. Eficiencia al transformar la masa en energía útil en relación con el límite teórico $E = mc^2$. Como se explica en el texto, el proceso de alimentar agujeros negros y esperar a que se evaporen, cuya eficacia es del 90 %, desafortunadamente es demasiado lento para ser útil, y la aceleración del proceso reduce drásticamente su eficiencia.



FIGURA 6.3. Una tecnología avanzada puede extraer enormemente más energía de la materia de la que obtenemos al comerla o quemarla, e incluso la fusión nuclear solo permite extraer una energía 140 veces menor que los límites que establecen las leyes físicas. Si existiesen centrales eléctricas que hiciesen uso de esfalerones, cuásares o de la evaporación de agujeros negros, obtendrían una eficiencia mucho mayor.

Evaporación de agujeros negros

En su libro *Una breve historia del tiempo*, Stephen Hawking propuso una central eléctrica de agujero negro.(19) Esto puede parecer paradójico, dado que durante mucho tiempo se creyó que los agujeros negros eran trampas de las que nada, ni siquiera la luz, podría escapar. Sin embargo, Hawking

calculó que los efectos de la gravedad cuántica hacen que un agujero negro actúe como un objeto caliente —cuanto más pequeño, más caliente— que emite radiación térmica, que ahora se conoce como radiación de Hawking. Esto significa que el agujero negro pierde energía gradualmente y se evapora. En otras palabras, cualquier materia que se arroje al agujero negro al final volverá a salir como radiación térmica, de manera que, cuando el agujero negro se haya evaporado por completo, habrá convertido su materia en radiación con casi un 100 % de eficiencia.(20)

Un problema de usar la evaporación del agujero negro como fuente de energía es que, a menos que el agujero negro sea mucho más pequeño que un átomo, es un proceso terriblemente lento, más largo que la edad actual del universo, y que irradia menos energía que una vela. La potencia producida disminuye con el cuadrado del tamaño del agujero; por ello, los físicos Louis Crane y Shawn Westmoreland han propuesto usar un agujero negro aproximadamente mil veces más pequeño que un protón, tan pesado como el mayor buque que jamás ha surcado los mares.[82] Su motivación principal era usar el agujero negro como motor para propulsar una nave estelar (un tema al que volveremos más adelante), por lo que les preocupaba más la portabilidad que la eficiencia, y su propuesta incluía alimentar el agujero negro con luz láser, sin causar en absoluto la conversión de energía en materia. Incluso si se pudiese alimentar con materia en lugar de radiación, alcanzar una alta eficiencia parece difícil: para conseguir que los protones entrasen en un agujero negro de una milésima parte de su tamaño, tendrían que lanzarse hacia el agujero usando una máquina tan potente como el Gran Colisionador de Hadrones (LHC, por sus siglas en inglés), para incrementar su energía mc^2 con al menos mil veces más energía cinética (de movimiento). Dado que al menos el 10 % de esa energía cinética se perdería en forma de gravitones cuando el agujero negro se evaporase, estaríamos poniendo más energía en el agujero negro de la que podríamos extraer y usar, lo que resultaría en una eficiencia negativa. Algo que oscurece todavía más las perspectivas de una central eléctrica de agujero negro es que aún no disponemos de una teoría rigurosa de la gravedad cuántica sobre la que basar nuestros cálculos, aunque esta incertidumbre podría significar, por supuesto, que existen efectos útiles de la gravedad cuántica aún por descubrir.

Agujeros negros rotatorios

Afortunadamente, hay otras formas de utilizar los agujeros negros como centrales eléctricas en las que no interviene la gravedad cuántica u otra física poco conocida. Por ejemplo, muchos agujeros negros giran muy rápido, de forma que sus horizontes de eventos dan vueltas a velocidades cercanas a la de la luz, y esta energía de rotación podría extraerse. El horizonte de sucesos de un agujero negro es la región de la que ni siquiera la luz puede escapar, porque la atracción gravitatoria es tan intensa que no lo permite. La figura 6.4 ilustra cómo, fuera del horizonte de sucesos, existe una región llamada *ergosfera*, donde el agujero negro rotatorio arrastra el espacio tan rápidamente que es imposible que una partícula se quede quieta y no sea arrastrada a su vez. Por lo tanto, si arrojásemos un objeto a la ergosfera, este ganará velocidad en su órbita alrededor del agujero. Por desgracia, enseguida será engullido por el agujero negro, desapareciendo para siempre en el horizonte de sucesos, por lo que no servirá de nada si lo que queremos es extraer energía. Sin embargo, Roger Penrose descubrió que si lanzamos el objeto en un ángulo concreto y hacemos que se divida en dos partes, como se muestra en la figura 6.4, podemos hacer que una sola parte sea engullida, mientras la otra escapa del agujero negro con más energía de la que teníamos al principio. Dicho de otro modo, habremos convertido parte de la energía de rotación del agujero negro en energía que podemos utilizar. Si repetimos este proceso muchas veces, podemos extraer del agujero negro toda su energía de rotación hasta que deje de girar y su ergosfera desaparezca. Si el agujero negro inicial girase tan rápido como la naturaleza lo permite, con su horizonte de eventos moviéndose básicamente a la velocidad de la luz, esta estrategia permitiría convertir el 29 % de su masa en energía. Todavía existe una gran incertidumbre acerca de la velocidad a la que giran los agujeros negros en el cosmos, pero muchos de los más estudiados parece que lo hacen bastante rápido: entre el 30 % y el 100 % de la velocidad máxima posible. Parece que el enorme agujero negro que existe en el centro de nuestra galaxia (que pesa cuatro millones de veces más que el Sol) es rotatorio, por lo que, incluso si únicamente el 10 % de su masa se pudiera convertir en energía útil, el resultado sería equivalente a la conversión en energía de 400.000 soles con una eficiencia del 100 %, casi tanta energía como la que obtendríamos de

esferas de Dyson construidas alrededor de 500 millones de soles durante miles de millones de años.

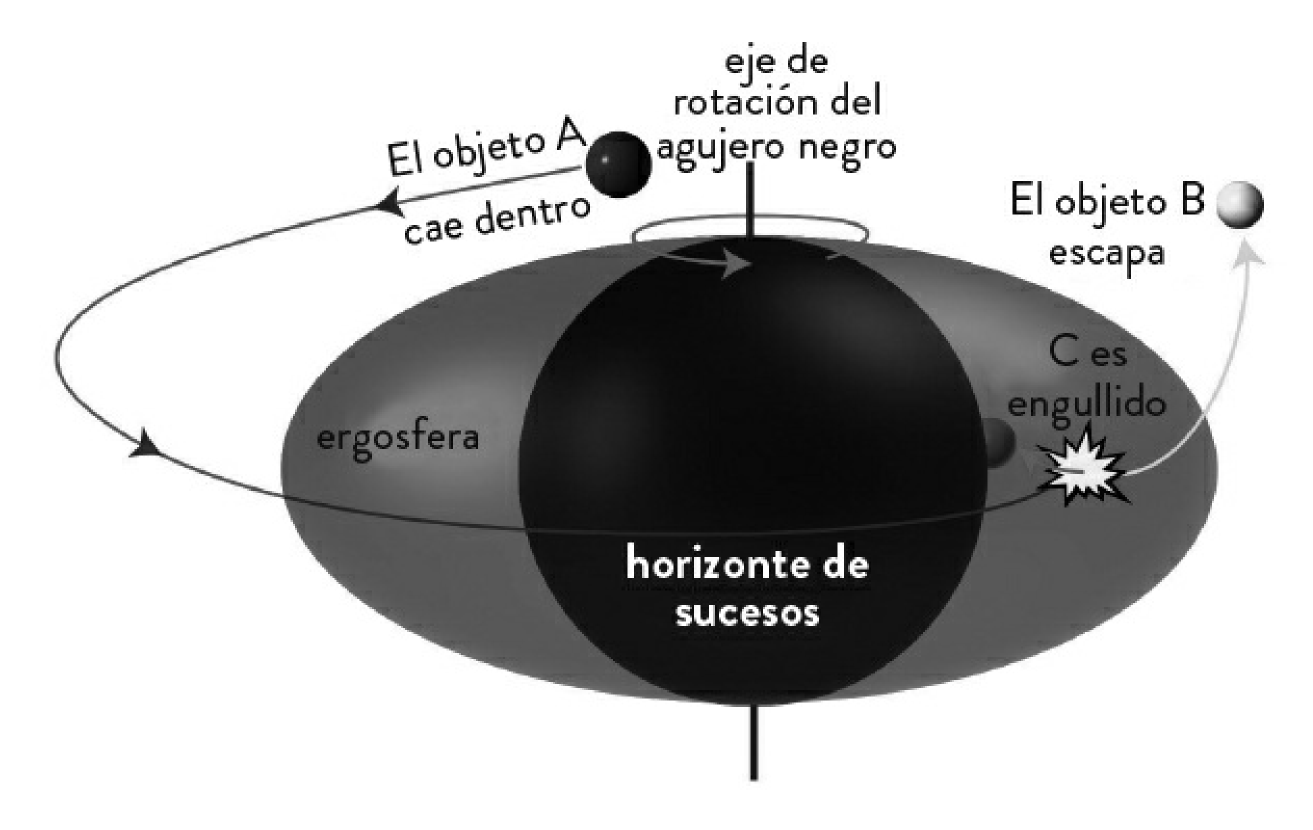


FIGURA 6.4. Se puede extraer parte de la energía de rotación de un agujero negro rotatorio arrojando una partícula A cerca del agujero negro y dividiéndola en una parte C que es engullida y una parte B que escapa con más energía de la que A tenía inicialmente.

Cuásares

Otra estrategia interesante consiste en extraer energía no del agujero negro en sí, sino de la materia que cae hacia él. La naturaleza ya ha encontrado por su cuenta una manera de hacer esto mismo: el cuásar. El gas que da vueltas cada vez más cerca de un agujero negro, formando un disco en forma de pizza cuyas zonas más internas son gradualmente engullidas, se calienta hasta alcanzar temperaturas elevadísimas y emite abundantes cantidades de radiación. A medida que el gas cae hacia el agujero, se acelera y transforma su energía potencial gravitatoria en energía cinética, como quien se lanza en paracaídas. El movimiento se vuelve cada vez más caótico y complicado a medida que las turbulencias convierten el desplazamiento coordinado de toda

la masa de gas en un movimiento aleatorio a escalas cada vez más pequeñas, hasta que los átomos individuales empiezan a chocar entre sí a elevadas velocidades (la existencia de este movimiento aleatorio es precisamente lo que significa que algo esté caliente, y estas violentas colisiones transforman la energía cinética en radiación). Mediante la construcción de una esfera de Dyson alrededor de todo un agujero negro, a una distancia segura, esta energía de radiación podría captarse y utilizarse. Cuanto más rápido gira un agujero negro, más eficiente es este proceso, hasta el punto de que un agujero negro que gira a la velocidad máxima posible permite obtener una eficiencia energética nada menos que del 42 %.(21) Para agujeros negros que pesen aproximadamente lo que una estrella, la mayor parte de la energía se emite en forma de rayos X, mientras que para los de tipo supermasivo que existen en los centros de las galaxias, buena parte se irradia en la franja de la luz infrarroja, visible y ultravioleta.

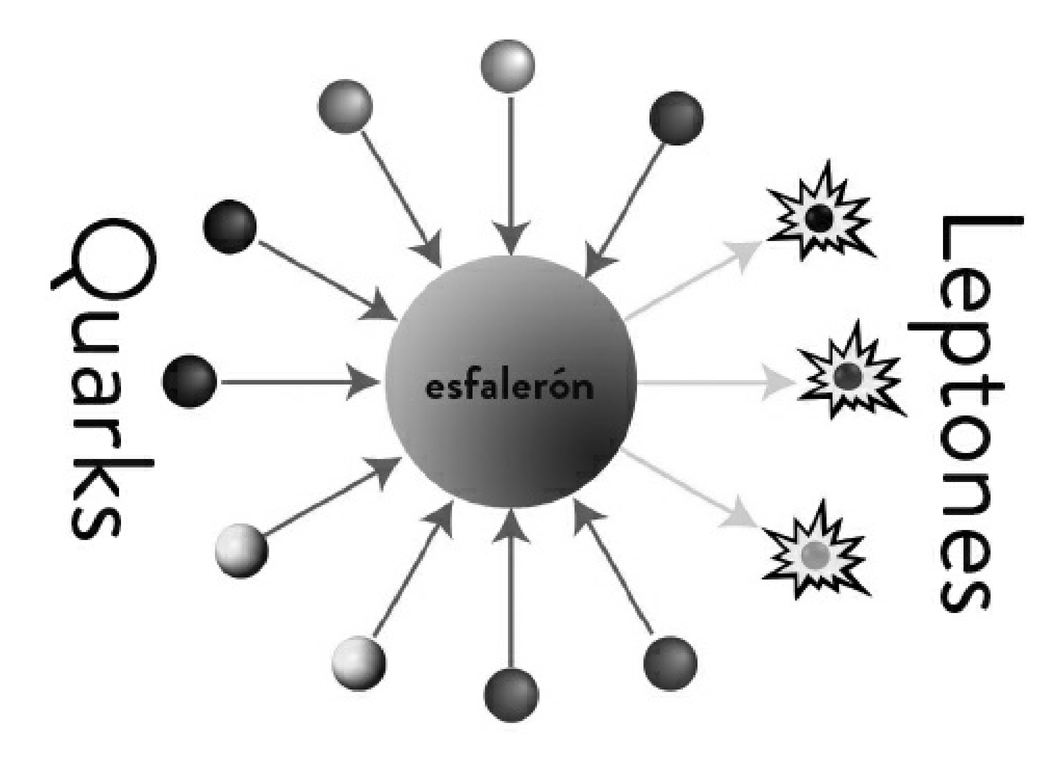


FIGURA 6.5. De acuerdo con el modelo estándar de la física de partículas, nueve quarks de sabor y espín apropiados pueden unirse y transformarse en tres leptones a través de un estado intermedio llamado esfalerón. La masa total de los quarks (junto con la energía de los gluones que los acompañaron) es mucho mayor que la masa de los leptones, por lo que este proceso liberará energía, lo que se representa en la figura mediante destellos.

Una vez que nos quedemos sin combustible con el que alimentar el agujero

negro, podemos pasar a extraer su energía rotacional, como vimos antes.(22) De hecho, la naturaleza también ha encontrado una manera de hacerlo parcialmente, potenciando la radiación del gas acumulado mediante un proceso magnético conocido como mecanismo de Blandford-Znajek. Quizá sea posible utilizar la tecnología para mejorar aún más la eficiencia de extracción de energía por encima del 42 % mediante el uso inteligente de campos magnéticos u otros elementos.

Esfalerones

Existe otra forma conocida de convertir la materia en energía en la que no interviene ningún agujero negro: el proceso del esfalerón. Dicho proceso puede destruir quarks y convertirlos en leptones: electrones; muones, sus primos más pesados; y partículas tau, neutrinos o sus antipartículas.[83] Como se ilustra en la figura 6.5, el modelo estándar de la física de partículas predice que nueve quarks de sabores y espines apropiados pueden juntarse y transformarse en tres leptones a través de un estado intermedio llamado esfalerón. Como el resultado final pesa más que todas las partículas iniciales juntas, la diferencia de masa se convierte en energía de acuerdo con la fórmula $E = mc^2$ de Einstein.

Por lo tanto, la vida inteligente del futuro podría ser capaz de construir lo que llamaré un *esfalerizador*: un generador de energía que funcionaría como un motor diésel elevado a la enésima potencia. Un motor diésel tradicional comprime una mezcla de aire y gasóleo hasta que la temperatura aumenta lo suficiente para que arda de forma espontánea, tras lo cual la mezcla caliente vuelve a expandirse, y al hacerlo realiza trabajo útil, por ejemplo, al empujar un pistón. El dióxido de carbono y otros gases de combustión pesan aproximadamente un 0,00000005 % menos que lo que el pistón contenía al principio, y esta diferencia de masa se transforma en la energía que impulsa el motor. Un esfalerizador comprimiría materia ordinaria hasta que esta alcanzase una temperatura de un par de miles de trillones de grados, y a continuación dejará que vuelva a expandirse y enfriarse una vez que los esfalerones han hecho su trabajo.(23) Ya sabemos el resultado de este experimento, porque en sus comienzos, hace unos 13.800 millones de años, el

universo lo llevó a cabo por nosotros cuando estaba tan caliente: casi el 100 % de la materia se convierte en energía, y menos de una mil millonésima parte de las partículas siguen existiendo para formar los componentes de la materia ordinaria: quarks y electrones. Es como un motor diésel, pero ¡mil millones de veces más eficiente! Otra ventaja es que no habría por qué ser muy cuidadoso con el combustible, ya que valdría cualquier cosa compuesta de quarks; esto es, cualquier materia normal.

Debido a estos procesos a altas temperaturas, el universo recién nacido produjo más de un billón de veces más radiación (fotones y neutrinos) que materia (quarks y electrones que posteriormente se apelotonaron en átomos). En los 13.800 millones de años transcurridos desde entonces, se produjo una gran segregación, en la cual los átomos se concentraron en galaxias, estrellas y planetas, mientras que la mayoría de los fotones permanecieron en el espacio intergaláctico, formando la radiación de fondo de microondas que se ha usado para tomar fotografías del universo en su más tierna infancia. Cualquier forma de vida avanzada que viva en una galaxia o en otra concentración de materia puede por lo tanto convertir la mayor parte de la materia a su alcance de nuevo en energía, devolviendo así el porcentaje de materia al mismo valor minúsculo que surgió del universo primitivo al recrear esas condiciones de temperatura y densidad elevadas en un esfalerizador.

Para calcular cuál sería la eficiencia de un esfalerizador real, necesitamos determinar varios detalles prácticos clave: por ejemplo, cuál debe ser su tamaño para evitar que una proporción significativa de los fotones y los neutrinos se escape durante la fase de compresión. Lo que sí podemos afirmar con seguridad es que las perspectivas energéticas para el futuro de la vida son radicalmente mejores de lo que nuestra tecnología actual hace posible. Ni siquiera hemos logrado construir un reactor de fusión, pero la tecnología futura debería ser capaz de hacer las cosas diez o incluso cien veces mejor.

Construir mejores ordenadores

Si la eficiencia de digerir la cena es diez mil millones de veces inferior al límite físico de la eficiencia energética, ¿cuál es la eficiencia de los ordenadores actuales? Peor aún que la de esa digestión, como ahora veremos.

A menudo presento a mi amigo y colega Seth Lloyd como la única persona en el MIT que puede que esté tan loca como yo. Después de hacer un trabajo pionero en ordenadores cuánticos, escribió un libro argumentando que todo el universo es una computadora cuántica. Solemos tomar unas cervezas después del trabajo, y aún no he descubierto un tema sobre el que Seth no tenga algo interesante que decir. Por ejemplo, como mencioné en el capítulo 2, tiene mucho que decir sobre los límites últimos de la informática. En un famoso artículo de 2000, mostró que la velocidad de cálculo está limitada por la energía: realizar una operación lógica elemental en un tiempo T requiere una energía promedio de $E = h/4T$, donde h es la cantidad de física fundamental conocida como constante de Planck. Esto significa que un ordenador de 1 kg puede realizar como máximo 5×10^{50} operaciones por segundo (nada menos que 36 órdenes de magnitud más que el ordenador en el escribo estas palabras). Llegaremos allí en un par de siglos si la potencia de computación se sigue duplicando cada dos años, como vimos en el capítulo 2. Seth también demostró que una computadora de 1 kg puede almacenar como mucho 10^{31} bits, aproximadamente un trillón de veces más que mi ordenador portátil.

Seth es el primero en reconocer que alcanzar estos límites puede ser complicado, incluso para la vida superinteligente, ya que la memoria de ese «ordenador» óptimo de 1 kg se asemejaría a una explosión termonuclear o a una pequeña parte del Big Bang. Sin embargo, es optimista en cuanto a que los límites prácticos no están tan lejos de los límites últimos. De hecho, los prototipos de ordenador cuántico existentes ya han logrado miniaturizar su memoria, al almacenar un bit por átomo, y si eso se lograse escalar, se podrían almacenar unos 10^{25} bits/kg, un billón de veces mejor que mi portátil. Además, el uso de radiación electromagnética para transmitir información entre estos átomos permitiría realizar en torno a 5×10^{40} operaciones por segundo, 31 órdenes de magnitud más que mi CPU.

En resumen, el potencial de la vida futura para calcular y resolver las cosas es realmente alucinante: en términos de órdenes de magnitud, los mejores supercomputadores actuales están mucho más lejos de la computadora óptima de 1 kg que de los intermitentes de un coche, un dispositivo que tan solo almacena un bit de información que alterna entre encendido y apagado aproximadamente una vez por segundo.

Otros recursos

Desde un punto de vista físico, todo lo que la vida futura puede querer crear —desde hábitats y máquinas hasta nuevas formas de vida— consiste simplemente en partículas elementales dispuestas de alguna manera particular. Así como una ballena azul es krill reorganizado y el krill es plancton reorganizado, todo el sistema solar no es más que hidrógeno reorganizado a lo largo de 13.800 millones de años de evolución cósmica: la gravedad reorganizó el hidrógeno en estrellas, que reorganizaron el hidrógeno en átomos más pesados, después de lo cual la gravedad reorganizó esos átomos en nuestro planeta, donde los procesos químicos y biológicos los reordenaron en vida.

La vida futura que ha alcanzado su límite tecnológico puede realizar esas reorganizaciones de partículas de manera más rápida y eficiente, usando primero su potencia de computación para descubrir el método más eficiente, y a continuación la energía de que dispone para activar el proceso de reorganización de la materia. Vimos cómo la materia se puede convertir tanto en ordenadores como en energía, por lo que en cierto sentido es el único recurso fundamental que se necesita.(24) Una vez que la vida futura ha alcanzado los límites físicos de lo que puede hacer con su materia, solo queda una manera de hacer más: obtener más materia. Y la única forma de hacerlo es expandiéndose por el universo. ¡Hacia el espacio!

OBTENER RECURSOS A TRAVÉS DE LA COLONIZACIÓN CÓSMICA

¿Cuán grande es nuestra herencia cósmica? En particular, ¿qué límites superiores imponen las leyes de la física a la cantidad de materia que la vida puede utilizar? Nuestra herencia cósmica es asombrosamente grande, sin duda, pero ¿cómo de grande? La tabla 6.2 enumera algunos números clave. Hoy en día, nuestro planeta está 99,999999 % muerto en el sentido de que esta proporción de su materia no forma parte de la biosfera y no hace prácticamente nada útil para la vida, aparte de generar atracción gravitatoria y un campo magnético. Esto abre la posibilidad de que algún día se llegue a

usar una cantidad de materia cien millones de veces mayor para dar soporte activo a la vida. Si conseguimos hacer un uso óptimo de toda la materia del sistema solar (incluido el Sol), lo haremos otro millón de veces mejor. Colonizar la galaxia haría que nuestros recursos se multiplicasen otro billón de veces.

REGIÓN	PARTÍCULAS
Nuestra biosfera	10^{43}
Nuestro planeta	10^{51}
Nuestro sistema solar	10^{57}
Nuestra galaxia	10^{69}
Nuestro alcance si viajásemos a la mitad de la velocidad de la luz	10^{75}
Nuestro alcance si viajásemos a la velocidad de la luz	10^{76}
Nuestro universo	10^{78}

TABLA 6.2. Número aproximado de partículas de materia (protones y neutrones) que la vida futura puede aspirar a utilizar.

¿Hasta dónde se puede llegar?

Alguien podría pensar que podemos obtener recursos ilimitados colonizando tantas galaxias como queramos si tenemos paciencia suficiente para hacerlo, pero eso no es lo que la cosmología moderna sugiere. Sí, puede que el espacio en sí sea infinito, que contenga un número infinito de galaxias, estrellas y planetas; de hecho, eso es lo que predicen las versiones más sencillas de la *inflación*, el paradigma científico actualmente más difundido, para explicar lo que produjo el Big Bang hace 13.800 millones de años. Sin embargo, aunque haya infinitas galaxias, parece que solo podemos observar y alcanzar un número finito de ellas: podemos ver alrededor de 200.000 millones de galaxias, y colonizar no más de 10.000 millones.

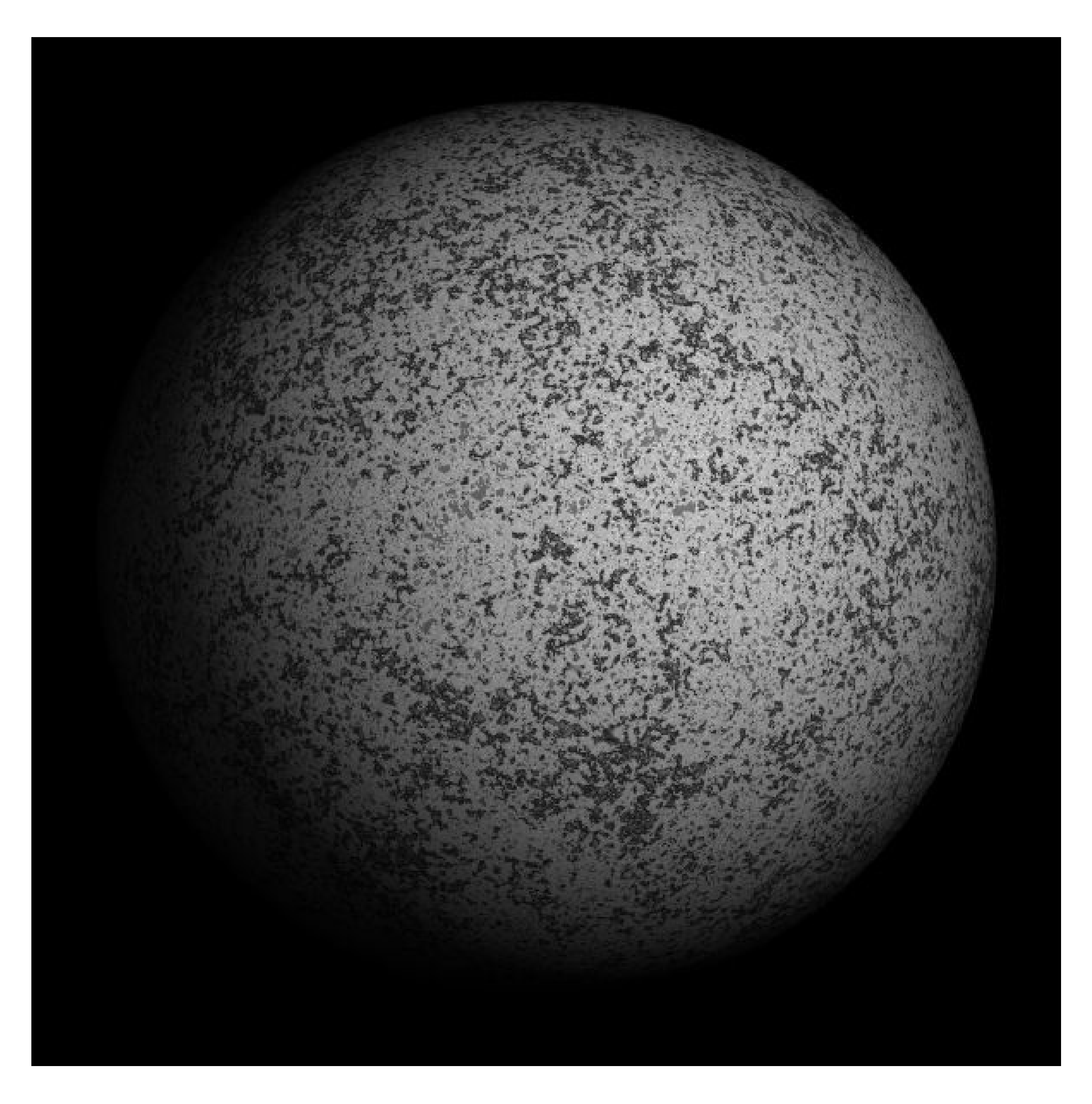


FIGURA 6.6. Nuestro universo, esto es, la región esférica del espacio desde la que la luz ha tenido tiempo de llegar hasta nosotros (en el centro) durante los 13.800 millones de años transcurridos desde el Big Bang. Las manchas muestran las fotos de nuestro universo cuando era bebé tomadas por el satélite Planck, en las que se observa que, cuando el universo tenía solo 400.000 años de antigüedad, estaba compuesto de plasma caliente casi tan caliente como la superficie del Sol. Es probable que el espacio continúe más allá de esta región, y con cada año que pasa se puede observar materia nueva.

Lo que nos limita es la velocidad de la luz: un año-luz (alrededor de diez billones de kilómetros) por año. La figura 6.6 muestra la parte del espacio desde la cual nos ha llegado luz hasta ahora durante los 13.800 millones de

años desde el Big Bang, una región esférica conocida como «el universo observable», o simplemente «nuestro universo». Incluso si el espacio es infinito, nuestro universo es finito, y contiene «solo» alrededor de 10^{78} átomos. Además, en torno al 98 % de nuestro universo «se ve pero no se toca», en el sentido de que podemos verlo pero nunca llegar a él, ni aunque viajásemos eternamente a la velocidad de la luz. ¿Por qué es esto? Al fin y al cabo, el límite de hasta dónde podemos ver viene dado por el hecho de que el universo no es infinitamente antiguo, de manera que la luz de procedencia remota aún no ha tenido tiempo de llegar hasta nosotros. ¿No deberíamos poder viajar a galaxias situadas a distancias arbitrariamente grandes si no existe un límite para el tiempo que podemos tardar en hacerlo?

La primera dificultad es que el universo se está expandiendo, lo que significa que casi todas las galaxias se están alejando de nosotros, por lo que colonizar galaxias remotas se convierte en un juego de «atrápame si puedes». La segunda complicación es que esta expansión cósmica está acelerándose, debido a la misteriosa energía oscura que forma aproximadamente el 70 % del universo. Para entender por qué esto es un problema, imaginemos que entramos en un andén y vemos cómo nuestro tren se aleja acelerando despacio, pero con una puerta abierta que invita a subirse. Si somos rápidos e imprudentes, ¿seríamos capaces de subirnos al tren? Puesto que su velocidad final será mayor que la que podamos alcanzar corriendo, la respuesta depende claramente de cuál sea la distancia que nos separa en un principio del tren: si supera determinada distancia crítica, nunca lograremos darle alcance. Una situación similar es la que se da al tratar de alcanzar esas galaxias remotas que se alejan de nosotros de forma acelerada: aunque pudiésemos viajar a la velocidad de la luz, todas las galaxias situadas a más de 17.000 millones de años luz de nosotros —el 98 % de las galaxias del universo— estarán para siempre fuera de nuestro alcance.

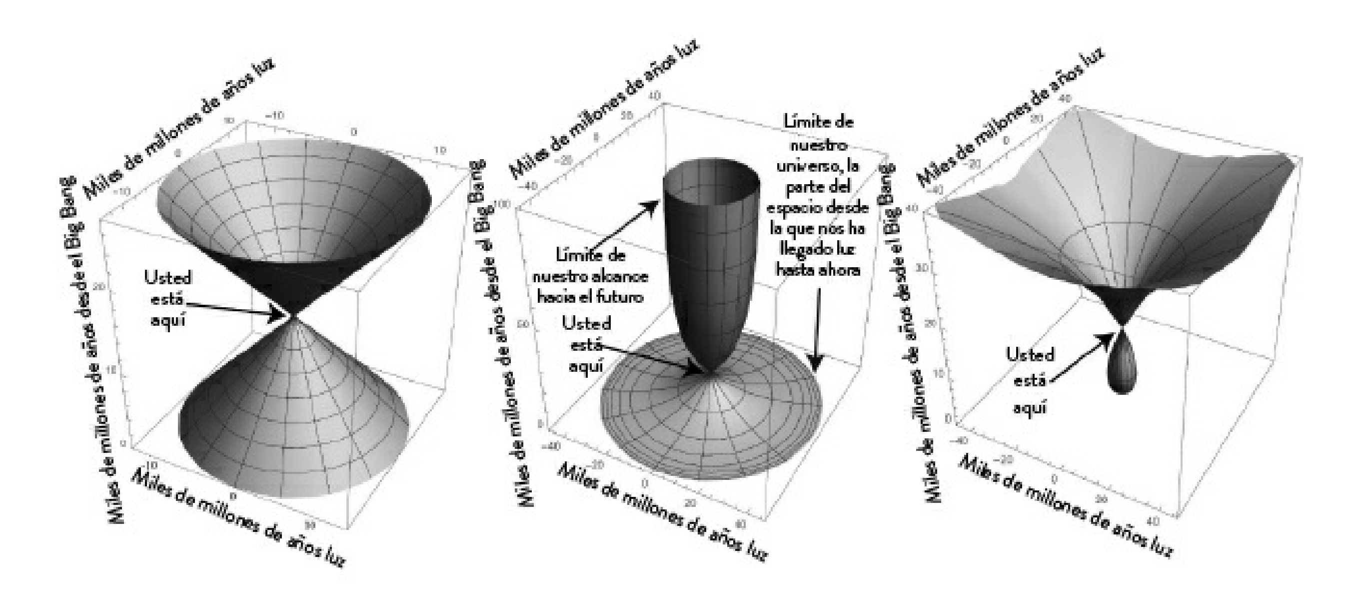


FIGURA 6.7. En un diagrama espaciotemporal, un evento es un punto cuyas posiciones horizontal y vertical marcan, respectivamente, dónde y cuándo ocurre. Si el espacio no se expande (imagen de la izquierda), los dos conos delimitan las partes del espacio-tiempo por las que nosotros en la Tierra (en el vértice) podemos vernos afectados (cono inferior) o sobre la que podemos tener algún efecto (cono superior), porque los efectos causales no pueden viajar más rápido que la luz, que recorre un año luz por año. Las cosas se ponen más interesantes cuando el espacio se expande (imagen central y de la derecha). Según el modelo estándar de la cosmología, incluso si el espacio es infinito, solo podemos ver y alcanzar una parte finita del espacio-tiempo. En la imagen del medio, que recuerda a una copa de champán, usamos coordenadas que ocultan la expansión del universo, de manera que los movimientos de galaxias remotas con el tiempo corresponden a líneas verticales. Desde nuestro punto de vista actual, 13.800 millones de años después del Big Bang, solo han tenido tiempo de llegarnos los rayos de luz procedentes de la base de la copa de champán, e incluso si viajásemos a la velocidad de la luz, nunca podríamos alcanzar las regiones situadas fuera de la parte superior de la copa, que contiene alrededor de 10.000 millones de galaxias. En la imagen de la derecha, que recuerda a una gota de agua debajo de una flor, usamos las coordenadas habituales en las que sí se ve cómo se expande el espacio. Esto deforma la base del espacio hasta que tiene forma de gota, porque las regiones en los extremos de lo observable estaban muy próximas entre sí al inicio.

Pero, un momento, ¿acaso la teoría de la relatividad especial de Einstein no decía que nada puede ir más rápido que la luz? ¿Cómo pueden las galaxias alejarse de algo que viaja a la velocidad de la luz? La respuesta está en la teoría de la relatividad general también de Einstein, que reemplazó a la de la relatividad especial, y en la que el límite de velocidad no es tan estricto: nada puede viajar más rápido que la velocidad de la luz *a través del espacio*, pero el espacio puede expandirse tan rápido como quiera. Einstein también nos proporcionó una manera efectiva de visualizar estos límites de velocidad, al interpretar el tiempo como la cuarta dimensión del espacio-tiempo (véase la

figura 6.7, donde la representación es tridimensional porque he omitido una de las tres dimensiones del espacio). Si el espacio no estuviese expandiéndose, los rayos de luz formarían líneas con una inclinación de 45 grados a través del espacio-tiempo, de manera que las regiones que podemos ver y alcanzar desde aquí y ahora son conos. Mientras que nuestro cono de luz pasado estaría truncado por el Big Bang hace 13.800 millones de años, nuestro cono de luz futuro se expandiría indefinidamente, dándonos acceso a una herencia cósmica ilimitada. Por su parte, en la imagen central de la figura se puede ver cómo un universo en expansión con energía oscura (como parece ser este que habitamos) deforma nuestros conos de luz hasta darles forma de copa de champán, lo que limita para siempre el número de galaxias en que podríamos colonizar a unos 10.000 millones.

Si este límite le provoca claustrofobia cósmica, permítame levantarle el ánimo con una posible escapatoria: mi cálculo supone que la energía oscura permanece constante en el tiempo, en consonancia con lo que sugieren las mediciones más recientes. Sin embargo, aún no tenemos ni idea de qué es realmente la energía oscura, lo que deja abierto un resquicio a la posibilidad de que la energía oscura pudiera decaer con el tiempo (como lo hace la sustancia similar a la energía oscura cuya existencia se postula para explicar la inflación cósmica) y, si esto sucede, la aceleración daría paso a una *deceleración*, lo que podría permitir que formas de vida futuras colonizasen nuevas galaxias durante toda su existencia.

¿Cuán rápido se puede ir?

Hemos visto cuántas galaxias podría colonizar una civilización si se expandiese en todas las direcciones a la velocidad de la luz. La relatividad general dice que es imposible enviar cohetes por el espacio a la velocidad de la luz, porque para ello sería necesaria una energía infinita; así pues, ¿a qué velocidad pueden ir los cohetes en la práctica?(25)

El cohete New Horizons de la NASA batió un récord de velocidad cuando fue lanzado hacia Plutón en 2006 a una velocidad aproximada de 160.000 kilómetros por hora (45 kilómetros por segundo), y la Solar Probe Plus de la NASA, prevista para 2018, aspira a alcanzar una velocidad cuatro veces superior al caer muy cerca del Sol, pero no deja de ser menos de un

insignificante 0,1 % de la velocidad de la luz. La búsqueda de cohetes más rápidos y mejores ha seducido a algunas de las mentes más brillantes del siglo pasado, y existe una rica y fascinante literatura al respecto. ¿Por qué es tan difícil ir más rápido? Los dos problemas esenciales son que los cohetes convencionales gastan la mayor parte de su combustible en acelerar el propio combustible que llevan consigo, y que el combustible para cohetes actual es extraordinariamente ineficiente: la proporción de su masa que se transforma en energía no es muy superior al 0,00000005 % de eficiencia para la gasolina que se recoge en la tabla 6.1. Una mejora evidente consiste en utilizar un combustible más eficiente. Por ejemplo, Freeman Dyson y sus colegas trabajaron en el Proyecto Orión de la NASA, que pretendía hacer explotar unas 300.000 bombas nucleares a lo largo de diez días para alcanzar en torno al 3 % de la velocidad de la luz con una nave espacial lo suficientemente grande para transportar humanos a otro sistema solar en un viaje de un siglo de duración.[84] Otros han investigado la forma de usar antimateria como combustible, ya que al combinarla con materia ordinaria se libera energía con una eficiencia cercana al 100 %.

Otra idea popular consiste en construir un cohete que no necesite transportar su propio combustible. Por ejemplo, el espacio interestelar no es un vacío perfecto, sino que contiene algún que otro ion de hidrógeno (un protón solo: un átomo de hidrógeno que ha perdido su electrón). En 1960, esto hizo que al físico Robert Bussard se le ocurriese la idea en la que se basa lo que ahora se conoce como *colector de Bussard*: ir recogiendo esos iones mientras la nave se desplaza y usarlos como combustible para los cohetes en un reactor de fusión incorporado a bordo. Aunque trabajos recientes han puesto en duda que sea posible que llegue a funcionar en la práctica, hay otra idea para evitar llevar combustible que sí parece factible para una civilización avanzada y que aspire a explorar el espacio: la navegación por láser.

La figura 6.8 ilustra un ingenioso diseño de un cohete de navegación láser propuesto en 1984 por Robert Forward, el mismo físico que inventó los estátites que vimos cuando hablamos de la construcción de una esfera de Dyson. Igual que las moléculas de aire propulsan un velero al rebotar contra la vela, las partículas de luz (fotones) que rebotasen contra un espejo lo impulsarían hacia delante. Dirigiendo un enorme haz láser alimentado por energía solar contra una gigantesca vela ultraligera sujeta a una nave espacial, podemos usar la energía del propio Sol para acelerar el cohete hasta

velocidades muy elevadas. Pero ¿cómo se detiene la nave? Esta es la cuestión para la que yo no encontraba respuesta hasta que leí el excelente artículo de Forward: como se ve en la figura 6.8, el anillo exterior de la vela láser se separa, se coloca delante de la nave espacial, y refleja el haz láser para decelerar la nave y su vela ahora más pequeña.[85] Forward calculó que esto podría permitir a los humanos realizar el viaje de cuatro años luz hasta el sistema solar Alfa Centauri en tan solo cuarenta años. Una vez allí, se supone que construiríamos un nuevo sistema láser gigante e iríamos saltando de estrella en estrella por toda la Vía Láctea.

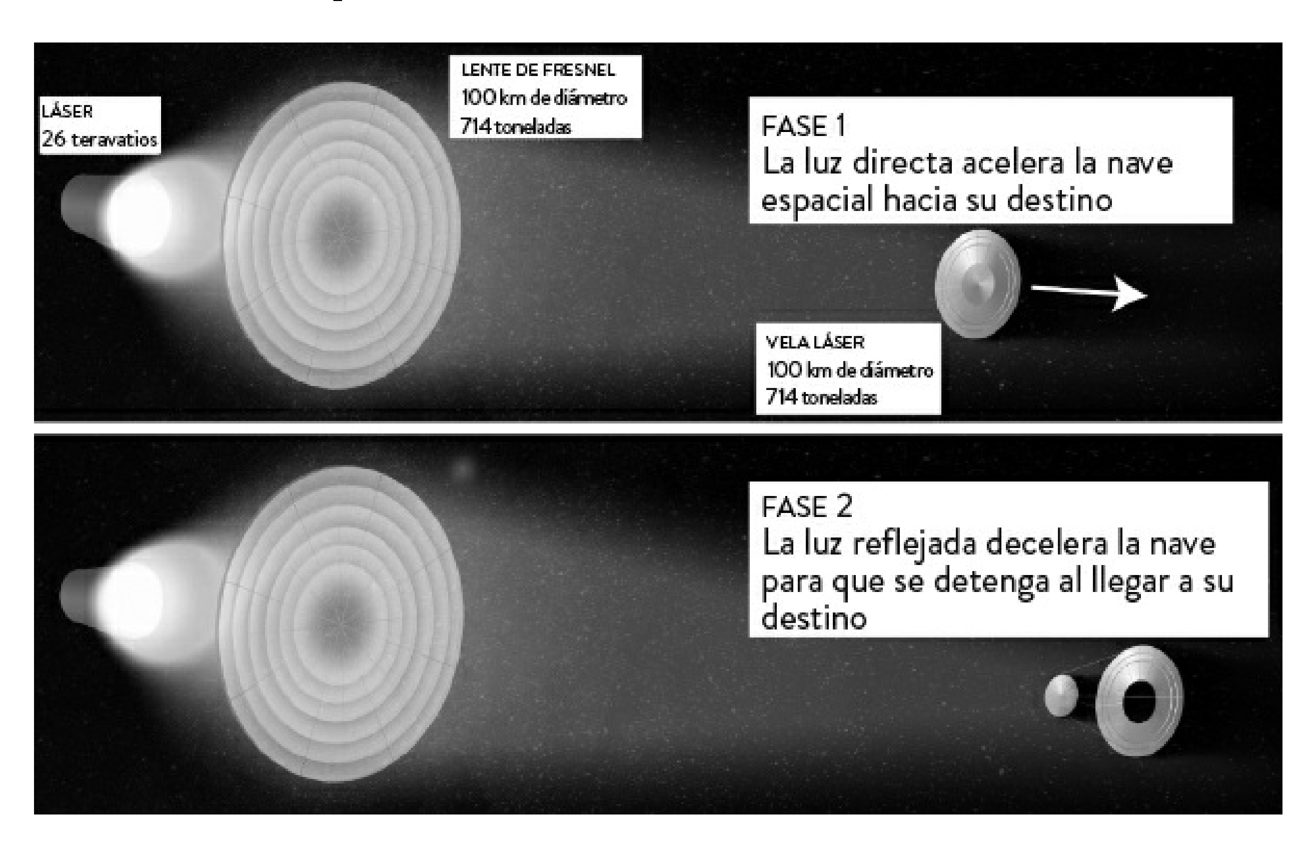


FIGURA 6.8. El diseño de Robert Forward para una misión de navegación láser al sistema estelar Alpha Centauri, situado a cuatro años luz de distancia. Inicialmente, en nuestro sistema solar, un potente láser acelera la nave espacial aplicando presión de radiación láser sobre su vela. Para frenar antes de llegar al destino, la parte exterior de la vela se separa y refleja la luz del láser sobre la nave espacial.

Pero ¿por qué detenerse ahí? En 1964, el astrónomo soviético Nikolái Kardashov propuso clasificar las civilizaciones en función de cuánta energía eran capaces de utilizar. Aprovechar la energía de un planeta, una estrella (con una esfera de Dyson, por ejemplo) y una galaxia corresponde a civilizaciones de tipo I, II y III en la escala de Kardashov, respectivamente.

Pensadores posteriores han propuesto que el tipo IV debería corresponder a las civilizaciones capaces de aprovechar la energía de todo el universo accesible. Desde entonces, ha habido noticias malas y noticias buenas para las formas de vida ambiciosas. La mala noticia es que existe la energía oscura, lo cual, como hemos visto, parece limitar nuestro alcance. La buena noticia es el espectacular progreso de la inteligencia artificial. Incluso los visionarios optimistas como Carl Sagan consideraban que las posibilidades de que los humanos llegasen a otras galaxias eran muy remotas, dada nuestra propensión a morir durante el primer siglo de un viaje que duraría millones de años, incluso aunque lo hiciésemos a velocidades próximas a la de la luz. Pero su negativa a darse por vencidos los llevaba a considerar la posibilidad de congelar a los astronautas para prolongar sus vidas, de ralentizar su envejecimiento al viajar a velocidades muy cercanas a la de la luz, o de enviar a una comunidad que estaría viajando durante decenas de miles de generaciones, más tiempo del que la humanidad lleva existiendo.

La posibilidad de la superinteligencia trastoca por completo la situación, y hace que sea mucho más prometedora para los apasionados de los viajes intergalácticos. Al eliminar la necesidad de transportar voluminosos sistemas de soporte vital para los humanos y sustituirlos por tecnologías inventadas por la IA, de pronto la colonización intergaláctica parece incluso algo sencillo. La navegación por láser de Forward se abarata mucho cuando la nave espacial solo necesita tener el tamaño suficiente para contener una «sonda seminal»: un robot capaz de aterrizar sobre un asteroide o planeta en el sistema solar elegido y construir allí una nueva civilización desde cero. Ni siquiera tiene que llevar consigo las instrucciones: lo único que necesita hacer es construir una antena lo suficientemente grande para recibir las instrucciones y los planos más detallados que su civilización nodriza le enviaría a la velocidad de la luz. Una vez hecho esto, usa sus láseres recién construidos para enviar nuevas sondas seminales a que continúen colonizando la galaxia de sistema solar en sistema solar. Incluso las oscuras inmensidades de espacio entre las galaxias suelen contener un número significativo de estrellas intergalácticas (desechos expulsados en algún momento de sus galaxias de origen) que podrían utilizarse como paradas intermedias, lo que permitiría ir de estrella en estrella para hacer posible la navegación por láser intergaláctica.

Una vez que la IA superinteligente hubiese colonizado otro sistema solar o

galaxia, llevar humanos allí sería fácil (si es que los humanos hubiesen conseguido que este fuese el objetivo de la IA). Toda la información necesaria sobre los humanos se puede transmitir a la velocidad de la luz, tras lo cual la IA puede juntar quarks y electrones para crear esos humanos. Esto podría hacerse bien usando tecnología relativamente poco avanzada, transmitiendo los dos gigabytes de información necesarios para especificar el ADN de una persona y a continuación incubando un bebé para que la IA lo criase, o bien esta podría nanoagrupar quarks y electrones para crear personas ya adultas, con todos sus recuerdos transferidos de los originales en la Tierra.

Esto significa que, si se produjese una explosión de inteligencia, la cuestión clave no es si la colonización intergaláctica es posible, sino simplemente a qué velocidad puede producirse. Puesto que todas las ideas que hemos explorado en los párrafos anteriores proceden de humanos, deben interpretarse como un límite inferior para la velocidad a la que la vida podría expandirse; a una vida superinteligente ambiciosa probablemente se le ocurrirían maneras mucho mejores de hacerlo, y además tendría un fuerte incentivo para ampliar sus horizontes, ya que, en la carrera contra el tiempo y la energía oscura, cada mejora del 1 % en velocidad media de colonización se traduce en un incremento del 3 % en la cantidad de galaxias colonizadas.

Por ejemplo, si se tardan 20 años en recorrer 10 años luz hasta el sistema estelar más próximo usando un sistema de navegación por láser, y luego otros 10 años para colonizarlo y construir allí nuevos láseres y sondas seminales, la región colonizada del espacio será una esfera que crezca en todas direcciones, en promedio, a un tercio de la velocidad de la luz. En un espléndido y concienzudo análisis de la expansión cósmica de las civilizaciones publicado en 2014, el físico estadounidense Jay Olson consideró una alternativa de tecnología avanzada a la estrategia de ir saltando de estrella en estrella que implicaba dos tipos distintos de sonda: *sondas seminales* y *expandidoras*.[86] Las sondas seminales se pararían, aterrizarían en su destino y lo sembrarían de vida. Las expandidoras, por su parte, nunca se detendrían: irían recogiendo materia a lo largo de su vuelo, quizá usando alguna variante mejorada del colector que vimos antes, y la utilizarían como combustible y como materia prima con la cual construir otras sondas seminales y copias de sí mismas. Esta flota autorreplicante de expandidoras iría acelerando gradualmente para mantener una velocidad constante (por ejemplo, de la mitad de la velocidad de la luz) respecto a las galaxias cercanas, y se reproduciría con la frecuencia

suficiente para que la flota formase una corteza esférica creciente con un número constante de expandidoras por unidad de área de la corteza.

Por último, pero no por ello menos importante, está también la aproximación del avemaría tramposa, que haría posible una expansión todavía más rápida que la de todos los métodos anteriores: usar el truco del «mensaje cósmico fraudulento» de Hans Moravec que vimos en el capítulo 4. Emitiendo un mensaje que engañase a ingenuas civilizaciones recién evolucionadas para que construyesen una máquina superinteligente que se haría con su control, una civilización podría expandirse básicamente a la velocidad de la luz, la velocidad a la que sus seductores cantos de sirena se propagarían por el cosmos. Puesto que esta podría ser la única manera que tendrían las civilizaciones avanzadas de llegar a la mayoría de las galaxias situadas dentro de sus conos de luz futuros, y tienen pocos incentivos para no ponerla a prueba, haríamos bien en desconfiar de cualquier transmisión extraterrestre. En el libro de Carl Sagan *Contacto*, los terrícolas usamos planos de procedencia extraterrestre para crear una máquina que no comprendíamos. No es algo que yo recomiende hacer…

En resumen, la mayoría de los científicos y escritores de ciencia ficción que han reflexionado sobre la colonización cósmica han sido, en mi opinión, excesivamente pesimistas al ignorar la posibilidad de la superinteligencia: al limitar su atención a los viajeros humanos, han sobreestimado la dificultad de los viajes intergalácticos, y, al limitarse a la tecnología inventada por humanos, han sobreestimado el tiempo que se tardaría en acercarse a los límites físicos de lo posible.

Permanecer conectados por medio de ingeniería cósmica

Si la energía oscura continúa haciendo que las galaxias remotas se alejen entre sí cada vez más rápido, como sugieren los datos experimentales más recientes, esto supondrá un obstáculo importante para el futuro de la vida. Significa que, incluso si una civilización futura logra colonizar un millón de galaxias, a lo largo de decenas de miles de millones de años, la energía oscura fragmentará este imperio cósmico en miles de regiones distintas incapaces de comunicarse entre sí. Si la vida futura no hace nada para evitar esta fragmentación, los mayores bastiones de vida serán cúmulos formados

por unas mil galaxias, cuya gravedad conjunta sea lo suficientemente fuerte para contrarrestar el efecto de la energía oscura que intenta separarlas.

El que una civilización superinteligente quiera permanecer conectada supondría un fuerte incentivo para que llevase a cabo ingeniería cósmica a gran escala. ¿Qué cantidad de materia tendría tiempo para transportar hasta el mayor de sus supercúmulos antes de que la energía oscura provocase que este pasase a ser para siempre inalcanzable? Un método para desplazar una estrella a mucha distancia consiste en aproximar una tercera estrella a un sistema binario en el que dos estrellas orbitan de forma estable la una alrededor de la otra. Como sucede en las relaciones sentimentales, la aparición de un tercero en discordia puede desestabilizar la situación y llevar a que uno de los tres sea expulsado violentamente (en el caso estelar, a gran velocidad). Si una o varias de esas tres partes son agujeros negros, un trío de naturaleza tan volátil podría utilizarse para proyectar masa con la velocidad suficiente para que acabase a enorme distancia de la galaxia de origen. Por desgracia, esta técnica de los tres cuerpos, aplicada a estrellas, agujeros negros o galaxias, no parece que vaya a permitir desplazar más que una minúscula proporción de toda la masa de una civilización a las enormes distancias necesarias para burlar a la energía oscura.

Pero esto obviamente no significa que la vida superinteligente no pueda inventar métodos mejores, por ejemplo, convertir buena parte de la masa de las galaxias alejadas en naves espaciales capaces de viajar hasta el cúmulo matriz. Si se pudiese construir un esfalerizador, quizá podría usarse para convertir la materia en energía susceptible de ser enviada en forma de luz hacia el cúmulo matriz, donde se podría reconfigurar de nuevo en materia o bien utilizarse como fuente energética.

Lo más afortunado sería que se pudiesen construir agujeros de gusano transitables, lo que permitiría las comunicaciones y los viajes casi instantáneos entre los dos extremos del agujero de gusano, con independencia de lo alejados que estuviesen entre sí. Un agujero de gusano es un atajo a través del espacio-tiempo que permite viajar de A a B sin tener que recorrer el espacio que los separa. Aunque la teoría de la relatividad general de Einstein admite la existencia de agujeros de gusano estables, como los que aparecen en las películas *Contact* e *Interstellar*, tales agujeros requieren una extraña clase hipotética de materia con densidad negativa, cuya existencia podría depender de efectos poco conocidos de la gravedad cuántica. Dicho de

otro modo, puede que los agujeros de gusano útiles sean imposibles, pero, si no es así, la vida superinteligente tendría una enorme motivación para construirlos: no solo revolucionarían las comunicaciones a gran velocidad dentro de cada galaxia, sino que, al conectar desde el principio las galaxias distantes con el cúmulo matriz, los agujeros de gusano permitirían que toda la extensión de la vida futura continuase conectada a largo plazo, obstaculizando por completo los intentos de la energía oscura de impedir la comunicación. Una vez que dos galaxias estuviesen conectadas por un agujero de gusano estable, seguirían estándolo por mucho que se alejasen con el transcurso del tiempo.

Si, a pesar de todos sus esfuerzos en torno a la ingeniería cósmica, una civilización futura llega a la conclusión de que algunas de sus partes están abocadas a alejarse hasta perder contacto para siempre, podría limitarse a dejar que esto sucediera y desearles la mejor de las suertes. Sin embargo, si la civilización ambiciona dedicar su capacidad de computación a buscar las respuestas a determinadas preguntas muy difíciles, podría recurrir a una estrategia de tierra quemada: podría convertir las galaxias distantes en gigantescos ordenadores que transformasen su materia y su energía en computación a un ritmo enloquecido, confiando en que, antes de que la energía oscura alejase sus despojos ya consumidos hasta hacerlos desaparecer de la vista, pudiesen transmitir al cúmulo matriz esas respuestas tan buscadas. Esta estrategia de tierra quemada sería particularmente apropiada para regiones tan distantes que solo se pudiese acceder a ellas usando el método del «mensaje cósmico fraudulento», para gran desgracia de sus habitantes previos. Mientras tanto, en la región matriz, la civilización podría aspirar a la máxima conservación y eficiencia para perdurar el mayor tiempo posible.

¿Cuánto se puede perdurar?

La longevidad es algo a lo que aspiran la mayoría de las personas, organizaciones y países ambiciosos. Siendo así, si una civilización futura ambiciosa desarrolla superinteligencia y se propone perdurar, ¿cuánto tiempo podría hacerlo?

El primer análisis científico riguroso de nuestro futuro remoto lo llevó a cabo el mismísimo Freeman Dyson, y la tabla 6.3 resume algunos de sus

resultados más importantes. La conclusión es que, a menos que intervenga la inteligencia, los sistemas solares y las galaxias se irán destruyendo progresivamente, seguidos con el tiempo por todo lo demás, hasta que no quede más que espacio vacío, frío y muerto, con un resplandor de radiación que irá desvaneciéndose a lo largo de toda la eternidad. Pero Freeman termina su análisis con un toque de optimismo: «Existen buenas razones científicas para tomarse en serio la posibilidad de que la vida y la inteligencia logren moldear a sus propósitos este universo nuestro».[87]

Creo que la superinteligencia podría resolver fácilmente muchos de los problemas que se enumeran en la tabla 6.3, puesto que podría reorganizar la materia en algo mejor que sistemas solares y galaxias. Otras dificultades que se suelen citar, como la muerte del Sol en unos cuantos miles de millones de años, no serían obstáculos insalvables, ya que incluso una civilización que disponga de tecnología relativamente poco avanzada podría trasladarse con facilidad a estrellas poco masivas que durarán más de 200.000 millones de años. Suponiendo que las civilizaciones superinteligentes construyan sus propias centrales eléctricas más eficientes que las estrellas, podrían de hecho querer *impedir* la formación de estrellas para conservar energía: incluso si utilizasen una esfera de Dyson para recoger toda la energía emitida durante la fase principal de la vida de una estrella (recuperando alrededor del 0,1 % de la energía total), podrían no ser capaces de evitar que gran parte del 99,9 % restante se perdiese cuando mueren estrellas muy grandes. Una estrella pesada muere en una explosión de supernova, de la cual la mayor parte de la energía escapa en forma de casi indetectables neutrinos, y una gran proporción de la masa se desperdicia al formar un agujero negro, del cual la energía tarda en filtrarse 10^{67} años.

QUÉ	CUÁNDO
Edad actual del universo	10^{10} años
La energía oscura aleja la mayoría de las galaxias hasta ser inalcanzables	10^{11} años
Últimas estrellas	10^{14} años
Los planetas se separan de las estrellas	10^{15} años
Las estrellas se separan de las galaxias	10^{19} años
Desintegración de las órbitas por radiación gravitatoria	10^{20} años
Desintegración de los protones (como muy pronto)	

	> 10^{34} años
Se evaporan los agujeros negros de masa estelar	10^{67} años
Se evaporan los agujeros negros supermasivos	10^{91} años
Toda la materia se degrada en hierro	10^{1500} años
Toda la materia forma agujeros negros, que después se evaporan	$10^{10^{26}}$ años

TABLA 6.3. Estimaciones para el futuro remoto (todas ellas, salvo la segunda y la séptima, hechas por Freeman Dyson). Dyson hizo estos cálculos antes del descubrimiento de la energía oscura, que podría hacer posibles varios tipos de «cosmocalipsis» en 10^{10}–10^{11} años. Los protones podrían ser completamente estables; si no lo son, los experimentos sugieren que tendrán que pasar al menos 10^{34} años para que la mitad de ellos se desintegren.

Si la vida superinteligente no se queda sin materia/energía, podrá mantener su hábitat en el estado que desee. Quizá pueda incluso descubrir un modo de evitar que los protones se desintegren, usando el llamado *efecto de la olla observada* de la mecánica cuántica, según el cual el proceso de desintegración de un objeto se ralentiza si se realizan observaciones frecuentes del mismo. Sí existe, no obstante, un posible fallo definitivo: un *cosmocalipsis* que destruya el universo, quizá dentro de solo entre 10.000 y 100.000 millones de años. El descubrimiento de la energía oscura y los avances en teoría de cuerdas han planteado nuevos escenarios de cosmocalipsis que Freeman Dyson desconocía cuando escribió su seminal artículo.

¿Cómo terminará el universo, dentro de miles de millones de años? Tengo cinco sospechosos principales como potenciales autores del apocalipsis cósmico, o cosmocalipsis, que se ilustran en la figura 6.9: el *Big Chill*, el *Big Crunch*, el *Big Rip*, el *Big Snap* y las *Burbujas letales*. El universo se ha expandido durante unos 13.800 millones de años. El Big Chill se produce cuando el universo se expande para siempre, diluyendo nuestro cosmos en un lugar frío, oscuro y, al final, muerto (esta era la situación que se consideraba más probable cuando Freeman escribió ese documento). Yo la veo como la opción de T. S. Eliot: «Así es como termina el mundo / no con una explosión, sino con un lamento». Si usted, como Robert Frost, prefiere que todo termine convertido en fuego en lugar de hielo, entonces cruce los dedos para que llegue el Big Crunch, donde la expansión cósmica se invierte y todo choca de nuevo en un colapso cataclísmico similar a un Big Bang al revés. Por último,

el Big Rip es como el Big Chill para impacientes, y en él las galaxias, los planetas e incluso los átomos se desgarran en un gran final transcurrido un tiempo finito desde ahora. ¿Por cuál de estos tres apostaría? Eso depende de lo que haga la energía oscura, que compone aproximadamente el 70 % de la masa del universo, a medida que el espacio continúe expandiéndose. Puede darse cualquiera de los escenarios (Chill, Crunch o Rip), en función de si la energía oscura se mantiene sin cambios, se diluye hasta alcanzar una densidad negativa, o bien se «antidiluye» hacia una mayor densidad, respectivamente. Como todavía no tenemos ni idea de qué es la energía oscura, esta sería mi apuesta: un 40 % al Big Chill, un 9 % al Big Crunch y un 1 % al Big Rip.

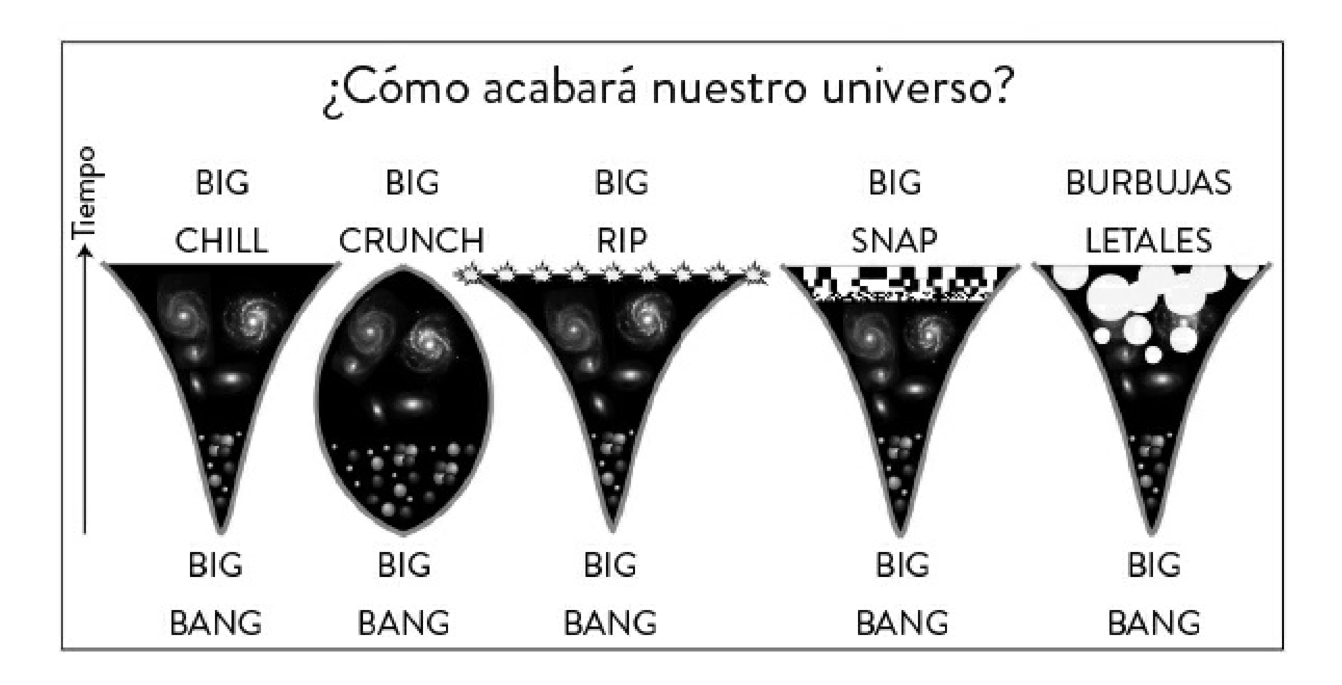


FIGURA 6.9. Sabemos que el universo comenzó con un caliente Big Bang hace 13.800 millones de años, se expandió y se enfrió, y fusionó sus partículas en átomos, estrellas y galaxias. Pero no sabemos su destino final. Los escenarios propuestos incluyen *Big Chill* (expansión eterna), *Big Crunch* (recolapso), *Big Rip* (una velocidad de expansión infinita que lo despedaza todo), *Big Snap* (la estructura del espacio revela una naturaleza granular letal cuando se estira demasiado) y *Burbujas letales* («congelación» del espacio en burbujas letales que se expanden a la velocidad de la luz).

¿Qué hay del otro 50 % de mi dinero? Lo reservo para la opción «ninguno de los anteriores», porque creo que los humanos debemos ser humildes y reconocer que hay cosas básicas que aún no comprendemos. La naturaleza del espacio, por ejemplo. Los finales en Chill, Crunch y Rip presuponen que

el espacio en sí es estable y puede estirarse indefinidamente. En épocas pasadas, pensábamos que el espacio era el aburrido escenario estático en el que se desarrolla el drama cósmico. Entonces Einstein nos enseñó que el espacio es uno de los actores clave: puede curvarse en agujeros negros, puede ondularse en forma de ondas gravitatorias y puede estirarse como un universo en expansión. Tal vez incluso se puede congelar y pasar a otra fase distinta, como hace el agua, y dentro de esa nueva fase podrían crecer rápidamente burbujas letales, que son otro posible candidato a causante del cosmocalipsis. Si fueran posibles las burbujas letales, es probable que se expandiesen a la velocidad de la luz, como la esfera creciente de difusión del mensaje cósmico fraudulento de una civilización muy agresiva.

Además, la teoría de Einstein dice que el estiramiento del espacio siempre puede continuar, permitiendo que el volumen del universo tienda a infinito, como en los escenarios del Big Chill y el Big Rip. Esto parece demasiado bueno para ser verdad, y sospecho que lo es. Una tira de goma parece continua, al igual que el espacio, pero si la estiramos demasiado se rompe. (26) ¿Por qué? Porque está hecha de átomos y, si se estira lo suficiente, esta naturaleza granular y atómica se vuelve relevante. ¿Podría ser que el espacio tenga algún tipo de granularidad a una escala demasiado pequeña para que la hayamos percibido? La investigación en gravedad cuántica sugiere que no tiene sentido hablar de espacio tridimensional tradicional a escalas inferiores a 10^{-34} metros. Si es realmente cierto que el espacio no puede estirarse indefinidamente sin sufrir un cataclísmico «Big Snap», las civilizaciones futuras tal vez opten por trasladarse a la mayor región del espacio que no esté en expansión (un gran cúmulo de galaxias) a la que puedan llegar.

¿Cuánto se puede computar?

Tras explorar cuánto *puede* durar la vida futura, veamos ahora cuánto podría *querer* durar. Aunque tal vez nos parezca natural querer vivir tanto como se pueda, Freeman Dyson también propuso un argumento más cuantitativo para explicar este anhelo: el coste de computación se reduce cuando la computación se realiza despacio, por lo que, en última instancia, se consigue hacer más cosas si se ralentiza el proceso lo máximo posible. Freeman

incluso llegó a demostrar con sus cálculos que, si el universo se sigue expandiendo y enfriando eternamente, se podría llegar a realizar una cantidad infinita de cálculos.

Lento no siempre significa aburrido: si la vida futura vive en un mundo simulado, el flujo temporal que experimente subjetivamente no tiene que parecerse en nada a la velocidad lentísima a la que la simulación se está ejecutando en el mundo exterior, por lo tanto la perspectiva de una computación infinita podría traducirse en inmortalidad subjetiva para las formas de vida simuladas. El cosmólogo Frank Tipler ha partido de esta idea para imaginar que también se podría alcanzar la inmortalidad subjetiva en los instantes finales antes de un Big Crunch si se acelerase la computación hacia el infinito a medida que la temperatura y la densidad se disparan.

Puesto que la energía oscura parece estropear los sueños de computación infinita tanto de Freeman como de Frank, la superinteligencia futura podría preferir consumir sus reservas de energía relativamente rápido, para convertirlas en computaciones antes de toparse con problemas como los horizontes cósmicos y la desintegración de los protones. Si maximizar la computación total es el objetivo último, la mejor estrategia será una a medio camino entre demasiado lenta (para evitar los problemas ya mencionados) y demasiado rápida (que consumiría más energía de la necesaria por cada unidad de computación).

Combinando todo lo que hemos visto en este capítulo, llegamos a la conclusión de que centrales eléctricas y ordenadores de la máxima eficiencia permitirían a la vida superinteligente realizar una asombrosa cantidad de computación. Para alimentar nuestro cerebro de trece vatios durante cien años se necesita la energía contenida en medio miligramo de materia (menos de lo que pesa un grano de azúcar). El trabajo de Seth Lloyd sugiere que el consumo energético del cerebro podría ser mil billones de veces más eficiente, con lo que la energía procedente de ese grano de azúcar bastaría para generar una simulación de todas las vidas humanas vividas a lo largo de la historia, así como miles de veces más personas. Si toda la materia en el universo accesible pudiese usarse para simular personas, eso permitiría simular 10^{69} vidas (o lo que fuese que la IA superinteligente prefiriese hacer con su capacidad de computación). El número de vidas podría ser aún mayor si sus simulaciones se ejecutasen a menor velocidad.[88] En cambio, en su

libro *Superinteligencia*, Nick Bostrom estima que se podrían simular 10^{58} vidas con hipótesis más conservadoras en cuanto a la eficiencia energética. Estos números son enormes con independencia de las vueltas que les demos, y tenemos la responsabilidad de asegurarnos de que este potencial futuro para el florecimiento de la vida no se echa a perder. En palabras de Bostrom: «Si representamos toda la felicidad experimentada a lo largo de toda una de estas vidas mediante una sola lágrima de alegría, entonces la felicidad de estas almas podría llenar y rellenar los océanos terrestres cada segundo, y seguir haciéndolo durante cien trillones de milenios. Es importantísimo que nos aseguremos de que son realmente lágrimas de alegría».

JERARQUÍAS CÓSMICAS

La velocidad de la luz limita no solo la rapidez con la que puede propagarse la vida, sino también su propia naturaleza, ya que impone fuertes limitaciones sobre la comunicación, la consciencia y el control. Si buena parte del cosmos cobra vida, ¿cómo será esta vida?

Jerarquías de pensamiento

¿Alguna vez ha intentado sin éxito matar una mosca con la mano? La razón por la que la mosca puede reaccionar más rápido es que es más pequeña, por lo que la información tarda menos tiempo en viajar entre sus ojos, cerebro y músculos. El principio de «más grande = más lento» es válido no solo en biología, donde el límite de velocidad viene dado por lo rápido que las señales eléctricas pueden viajar por las neuronas, sino también para la futura vida cósmica si la información no puede viajar más rápido que la luz. Así pues, para un sistema inteligente de procesamiento de información, ser más grande tiene tanto ventajas como inconvenientes, lo que exige alcanzar un equilibrio interesante. Por una parte, ser más grande permite que contenga más partículas, lo que hace posibles pensamientos más complejos. Por otra, esto reduce la velocidad a la que el sistema puede tener pensamientos realmente globales, ya que la información relevante ahora tarda más en

propagarse hasta todas sus partes.

Si la vida se extiende por todo el cosmos, ¿qué forma elegirá: simple y rápida, o compleja y lenta? Mi predicción es que elegirá lo mismo que ha elegido la vida en la Tierra: ¡ambas! Los habitantes de la biosfera terrestre abarcan un asombroso rango de tamaños, desde las descomunales ballenas azules de doscientas toneladas hasta la minúscula bacteria *Pelagibacter*, de 10^{-16} kg, que se cree que constituye más biomasa que todos los peces del mundo juntos. Además, los organismos grandes, complejos y lentos a menudo mitigan su lentitud al contener módulos más pequeños que son simples y rápidos. Por ejemplo, el reflejo del parpadeo es muy rápido precisamente porque está implementado por un circuito pequeño y sencillo en el que no interviene la mayor parte del cerebro: si esa mosca que no consigue matar se dirige de forma accidental hacia su ojo, usted parpadeará en menos de una décima de segundo, mucho antes de que la información relevante haya tenido tiempo de propagarse por su cerebro y hacer que tome consciencia de lo que ha sucedido. Al organizar su procesamiento de información en una jerarquía de módulos, nuestra biosfera consigue lo mejor de ambos extremos, grande y pequeño, al tener tanto velocidad como complejidad. Los humanos ya estamos usando esta misma estrategia jerárquica para optimizar la computación paralela.

Puesto que la comunicación interna es lenta y costosa, imagino que la vida cósmica avanzada del futuro también lo será, de forma que las computaciones se lleven a cabo lo más localmente posible. Si una computación es lo bastante simple para efectuarla con un ordenador de 1 kg, es contraproducente distribuirla en un ordenador del tamaño de una galaxia, ya que esperar a que la información se transmita a la velocidad de la luz tras cada paso de la computación provocaría un absurdo retardo de 100.000 años por paso.

Cuánto —si es que algo— de este procesamiento futuro de la información será consciente en el sentido de involucrar una experiencia subjetiva es un tema controvertido y fascinante que abordaremos en el capítulo 8. Si la consciencia requiere que las diferentes partes del sistema sean capaces de comunicarse entre sí, entonces los pensamientos de sistemas más grandes son necesariamente más lentos. Mientras que usted o un futuro superordenador del tamaño de la Tierra puede tener muchos pensamientos por segundo, una mente del tamaño de una galaxia podría tener solo un pensamiento cada cien

mil años, y una mente cósmica de mil millones de años luz solo tendría tiempo para tener aproximadamente diez pensamientos en total antes de que la energía oscura la desgarrare en partes desconectadas. Por otra parte, estos pocos y hermosos pensamientos, y las experiencias que los acompañasen, podrían ser muy profundos.

Jerarquías de control

Si el pensamiento en sí está organizado en una jerarquía que abarca un amplio rango de escalas, ¿qué pasa con el poder? En el capítulo 4, vimos cómo las entidades inteligentes se organizan de manera natural en jerarquías de poder en equilibrio de Nash, en el cual cualquier entidad saldría perdiendo si modificase su estrategia. Cuanto mejor sea la tecnología de la comunicación y del transporte, más pueden crecer estas jerarquías. Si la superinteligencia se expande algún día a escalas cósmicas, ¿cómo será su jerarquía de poder? ¿Será anárquica y descentralizada, o muy autoritaria? ¿La cooperación se basará fundamentalmente en el beneficio mutuo, o en la coerción y las amenazas?

Para arrojar algo de luz sobre estas cuestiones, consideremos la zanahoria y el palo: ¿qué incentivos existen para la colaboración a escala cósmica, y qué amenazas podrían emplearse para imponerla?

Controlar con la zanahoria

En la Tierra, el intercambio ha sido el impulsor tradicional de la cooperación debido a que la dificultad relativa de producir distintas cosas varía según los lugares del planeta. Si extraer un kilogramo de plata cuesta 300 veces más que extraer un kilogramo de cobre en una región, pero solo 100 veces más en otra, ambas saldrán ganando al intercambiar 200 kg de cobre por 1 kg de plata. Análogamente, si la tecnología en una región está mucho más avanzada que en otra, ambas pueden salir beneficiadas si intercambian bienes de alta tecnología por materias primas.

Sin embargo, si la superinteligencia desarrolla tecnología capaz de reorganizar fácilmente las partículas elementales en cualquier forma de

materia que se desee, eliminará buena parte de la motivación de los intercambios a larga distancia. ¿Por qué molestarse en transportar plata entre sistemas solares distantes cuando es más sencillo y más rápido transmutar cobre en plata reorganizando sus partículas? ¿Por qué molestarse en transportar maquinaria de alta tecnología entre galaxias cuando tanto el conocimiento y las materias primas (cualquier tipo de materia valdría) existen en ambos lugares? Preveo que, en un universo rebosante de superinteligencia, casi la única mercancía que merecerá la pena transportar a largas distancias será la información. La única excepción podría ser la de la materia que se usaría en los proyectos de ingeniería cósmica (por ejemplo, para contrarrestar la ya mencionada tendencia destructiva de la energía oscura a desgarrar civilizaciones enteras). A diferencia de lo que sucede en el comercio humano tradicional, la materia se podría transportar en cualquier forma en bruto que resultase conveniente, quizá incluso como un haz de energía, puesto que la superinteligencia que la recibiese podría reorganizarla rápidamente para crear cualesquiera objetos que desease.

Si el hecho de compartir o intercambiar información se convierte en el principal impulsor de la cooperación cósmica, ¿qué tipos de información serían los que se manejarían? Cualquier conocimiento deseable será valioso si para generarlo se necesita un esfuerzo computacional intenso y prolongado. Por ejemplo, una superinteligencia podría querer obtener respuestas a difíciles cuestiones científicas sobre la naturaleza de la realidad física, complicadas cuestiones matemáticas sobre teoremas y algoritmos óptimos, y arduas cuestiones ingenieriles sobre cuál es la mejor manera de construir tecnologías espectaculares. Unas formas de vida hedonistas podrían querer un entretenimiento digital impresionante y experiencias simuladas, y el comercio cósmico podría acrecentar la demanda de alguna clase de criptomoneda cósmica, en la línea de los bitcoines.

Tales oportunidades de compartir podrían incentivar el flujo de información no solo entre entidades con un poder aproximadamente equivalente, sino también entre distintos niveles de las jerarquías de poder, por ejemplo entre nodos del tamaño de un sistema solar y un centro de operaciones galáctico, o entre nodos del tamaño de galaxias y un centro de operaciones cósmico. A los nodos podría interesarles hacer esto por el gusto de formar parte de algo más grande, por obtener respuestas y tecnologías que no podrían desarrollar por su cuenta y para que les proporcionasen defensa

contra amenazas externas. También podrían valorar la promesa de cuasi inmortalidad a través de las copias de respaldo de la información: al igual que muchos humanos encuentran consuelo en la creencia de que sus mentes vivirán eternamente una vez que sus cuerpos físicos hayan muerto, una IA avanzada podría valorar que su mente y su conocimiento perviviesen en un superordenador central, una vez que su hardware físico original hubiese agotado todas sus reservas de energía.

Por otra parte, el centro podría querer que sus nodos lo ayudasen con descomunales tareas de computación a largo plazo de cuyos resultados no sería urgente disponer, por lo que merecería la pena esperar miles o millones de años hasta obtenerlos. Como vimos antes, el centro también podría querer que sus nodos lo ayudasen a llevar a cabo enormes proyectos de ingeniería cósmica, como los pensados para contrarrestar los efectos destructivos de la energía oscura al acercar entre sí concentraciones galácticas de masa. Si los agujeros de gusano transitables fuesen posibles y construibles, una de las máximas prioridades de un centro de operaciones sería probablemente construir una red de dichos agujeros para combatir los efectos de la energía oscura y mantener su imperio conectado indefinidamente. Las cuestiones en torno a qué objetivos últimos podría tener una superinteligencia cósmica son fascinantes y controvertidas, como veremos en mayor profundidad en el capítulo 7.

Control con el palo

Los imperios terrestres suelen impulsar a sus subordinados a cooperar mediante el uso de la zanahoria y el palo. Aunque los súbditos del Imperio romano apreciaban la tecnología, la infraestructura y la defensa que se les ofrecían a cambio de su cooperación, también temían las inevitables consecuencias de rebelarse o no pagar impuestos. Debido al tiempo que se tardaba en enviar tropas desde Roma hasta las provincias exteriores, parte de la intimidación se delegaba en tropas locales y oficiales leales que tenían la potestad de infligir castigos casi instantáneos. Un centro superinteligente podría usar la estrategia análoga de desplegar una red de leales guardias por todo su imperio cósmico. Puesto que los súbditos superinteligentes podrían resistirse al control, la estrategia viable más sencilla podría consistir en

utilizar IA guardianas programadas para ser absolutamente leales gracias al hecho de ser relativamente tontas, que se limitarían a supervisar el cumplimiento de todas las normas y activarían automáticamente un dispositivo del Apocalipsis de no ser así.

Supongamos, por ejemplo, que la IA central dispone que una enana blanca se sitúe cerca de una civilización del tamaño del sistema solar que desea controlar. Una enana blanca es la cáscara quemada de una estrella no demasiado pesada. Está compuesta en buena medida de carbono, se asemeja a un diamante gigante en el firmamento, y es tan compacta que puede pesar más que el Sol aun siendo más pequeña que la Tierra. Como es bien sabido, el físico indio Subrahmanyan Chandrasekhar demostró que si se le iba añadiendo materia a la enana blanca hasta que superase el *límite de Chandrasekhar*, de unas 1,4 veces la masa del Sol, la estrella experimentaría una cataclísmica detonación termonuclear denominada supernova de tipo 1A. Si la IA central ha dispuesto cruelmente que esta enana blanca se encuentre muy cerca de su límite de Chandrasekhar, la IA guardiana podría ser efectiva aunque fuese muy tonta (de hecho, en gran medida porque es tan tonta): se podría programar para que se limitase a verificar que la civilización subyugada había entregado su cuota mensual de bitcoines cósmicos, demostraciones matemáticas o cualquier otro impuesto que se hubiese estipulado, y, si no lo hubiese hecho, echaría la suficiente cantidad de masa en la enana blanca para desencadenar la supernova y hacer que toda la región volase en pedazos.

Las civilizaciones del tamaño de galaxias podrían ser igualmente controlables colocando grandes cantidades de objetos compactos en órbitas ajustadas alrededor del agujero negro descomunal en el centro de la galaxia, y amenazando con convertir estas masas en gas, por ejemplo haciendo que chocasen entre sí. Este gas empezaría entonces a alimentar el agujero negro, transformándolo en un potente cuásar, lo que podría hacer que buena parte de la galaxia se volviese inhabitable.

En resumen, existen fuertes incentivos para que la vida futura coopere a través de distancias cósmicas, pero no está nada claro si esa cooperación se basará sobre todo en el beneficio mutuo o en brutales amenazas: los límites que impone la física aparentemente permiten ambos escenarios, por lo que el resultado dependerá de los objetivos y valores que acaben imponiéndose. En el capítulo 7 reflexionaremos sobre nuestra capacidad para influir sobre estos

objetivos y valores.

Cuando las civilizaciones chocan

Hasta ahora, solo hemos comentado escenarios en los que la vida se expande por el cosmos a partir de una única explosión de inteligencia. Pero ¿qué sucede si la vida surge de forma independiente en más de un lugar y se encuentran dos civilizaciones en proceso de expansión?

Si consideramos un sistema solar cualquiera, hay cierta probabilidad de que la vida surja en alguno de sus planetas, desarrolle tecnología avanzada y se expanda hacia el espacio. Es una posibilidad a tener en cuenta, puesto que en nuestro sistema solar ha surgido la vida tecnológica, y las leyes de la física permiten la colonización espacial. Si el espacio es lo suficientemente grande (de hecho, la teoría de la inflación cosmológica sugiere que es inmenso o infinito), habrá muchas de estas civilizaciones en expansión, como se ilustra en la figura 6.10. El artículo de Jay Olson ya mencionado incluye un elegante análisis de esas biosferas cósmicas en expansión, y Toby Ord ha llevado a cabo un análisis similar junto con varios colegas en el Future of Humanity Institute. Vistas en tres dimensiones, estas biosferas cósmicas son, en un sentido muy literal, esferas siempre que las civilizaciones se expandan a la misma velocidad en todas las direcciones. En el espacio-tiempo, su aspecto recuerda a la parte superior de la copa de champán de la figura 6.7, porque la energía oscura limita en última instancia a cuántas galaxias puede llegar cada civilización.

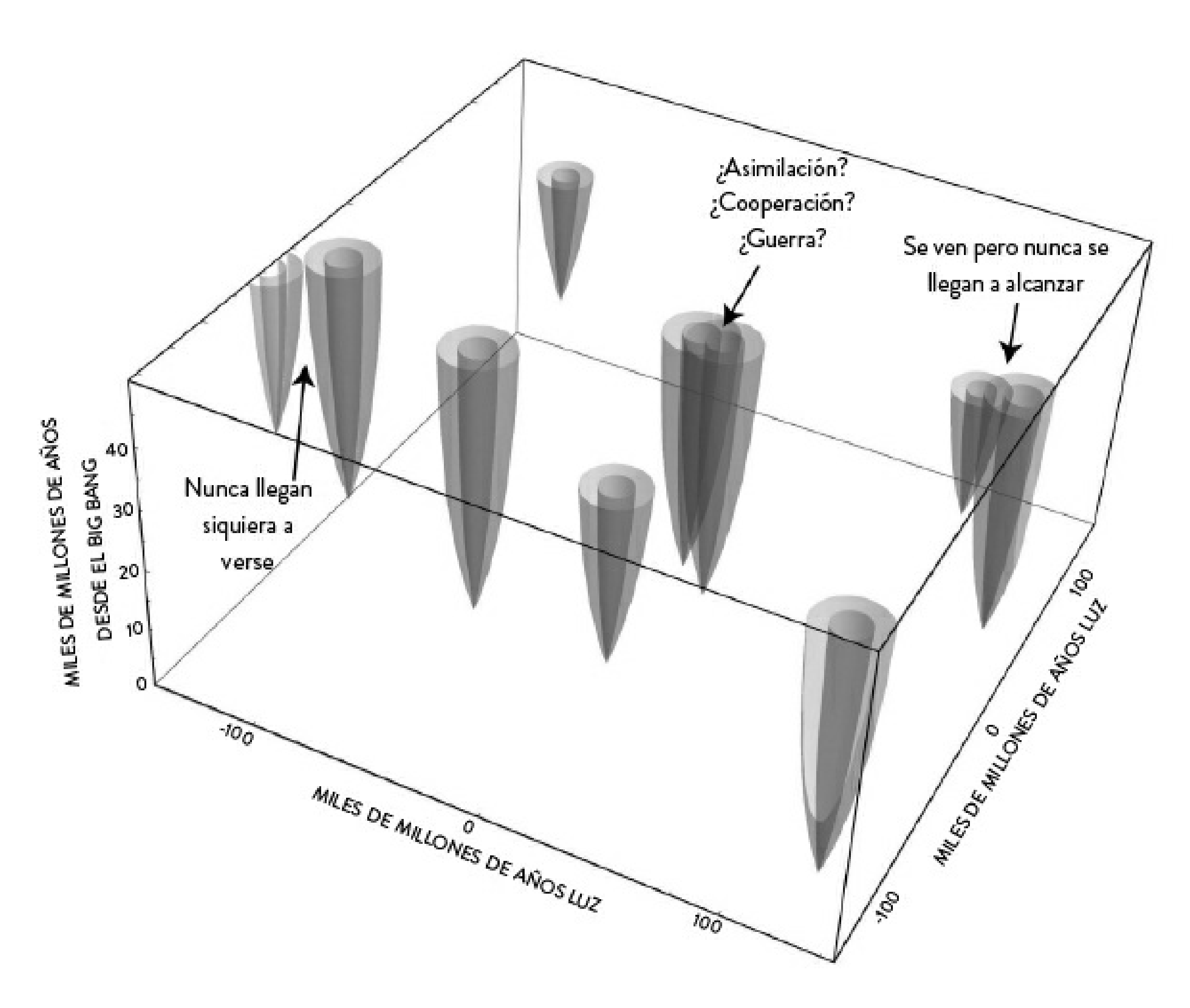


FIGURA 6.10. Si la vida surge de forma independiente en muchos puntos del espacio-tiempo (lugares y tiempos) y comienza a colonizar el espacio, entonces este contendrá una red de biosferas cósmicas en expansión, cada una de las cuales tiene un aspecto similar a la parte superior de la copa de champán de la figura 6.7. La parte inferior de cada biosfera representa el lugar y el momento en que comenzó la colonización. Las copas de champán opacas y translúcidas corresponden a la colonización al 50 % y al 100 % de la velocidad de la luz, respectivamente, y los solapamientos muestran dónde se encuentran las civilizaciones independientes.

Si la distancia entre civilizaciones colonizadoras del espacio es muy superior al límite hasta el que la energía oscura permite que se expandan, nunca entrarán en contacto ni sabrán siquiera de la existencia de las otras civilizaciones, por lo que creerán que están solas en el universo. Sin embargo, si el cosmos es más fecundo y las civilizaciones vecinas están más próximas entre sí, llegará un momento en que se solapen. ¿Qué sucede en estas regiones de solapamiento? ¿Habrá cooperación, competencia o guerra?

Los europeos fueron capaces de conquistar África y América porque eran

tecnológicamente superiores. Por otra parte, es posible que, mucho antes de que dos civilizaciones superinteligentes se encuentren, la evolución de sus tecnologías se estanque en el mismo nivel, limitada solo por las leyes de la física. Esto hace que parezca improbable que una superinteligencia pueda imponerse fácilmente sobre otra aunque lo desease. Además, si sus objetivos han evolucionado hasta estar más o menos alineados, podrían tener pocos motivos para desear la conquista o la guerra. Por ejemplo, si ambas están intentando demostrar el máximo número posible de hermosos teoremas e inventar los algoritmos más ingeniosos, podrían simplemente compartir lo que saben, de manera que las dos salieran ganando. Al fin y al cabo, la información es muy distinta de otros recursos por los que los humanos suelen pelearse, porque uno puede al mismo tiempo darla y conservarla.

Algunas de las civilizaciones en expansión podrían tener objetivos que fuesen esencialmente inmutables, como los de una secta fundamentalista o los que impulsan la propagación de un virus. Sin embargo, también es posible que algunas civilizaciones avanzadas sean más bien como las personas abiertas de mente: dispuestas a ajustar sus objetivos cuando se les presenten argumentos lo bastante convincentes. Si dos de ellas se encuentran, se producirá un choque no armado sino de ideas, en el que la civilización más persuasiva se impondrá y conseguirá que sus objetivos se propaguen a la velocidad de la luz a través de la región controlada por la otra civilización. Para una civilización, asimilar a sus vecinos es una estrategia de expansión más rápida que la colonización, puesto que la esfera de influencia de dicha civilización puede ampliarse a la velocidad con la que se transmiten las ideas (la velocidad de la luz, si se usan las telecomunicaciones), mientras que la colonización física avanza inevitablemente a una velocidad más lenta que la de la luz. Esta asimilación no se produciría por la fuerza, como la que emplean los Borg en *Star Trek*, sino que sería voluntaria, basada en la superior capacidad de persuasión de las ideas, que mejorarían la vida de aquellos que han sido asimilados.

Hemos visto que el cosmos del futuro puede contener burbujas en rápida expansión de dos tipos: civilizaciones en proceso de expansión y burbujas letales que crecen a la velocidad de la luz y hacen el espacio inhabitable al destruir todas nuestras partículas elementales. Así pues, una civilización ambiciosa puede encontrarse tres tipos de regiones: deshabitadas, burbujas de vida y burbujas de muerte. Si teme toparse con civilizaciones rivales poco

dispuestas a cooperar, tiene una fuerte motivación para lanzar una rápida «anexión de territorios» y colonizar las regiones deshabitadas antes de que lo hagan sus rivales. Pero esta misma motivación persiste aunque no haya civilizaciones rivales, simplemente para adquirir recursos antes de que la energía oscura haga que sean inalcanzables. Acabamos de ver cómo toparse con otra civilización en expansión puede ser mejor o peor que encontrar un espacio deshabitado, dependiendo de lo cooperativos y abiertos que sean esos vecinos. Sin embargo, es mejor encontrarse con cualquier civilización expansionista (incluso con una que intente convertir a la nuestra en clips) que con una burbuja letal, que seguirá expandiéndose a la velocidad de la luz con independencia de que intentemos razonar con ella o combatirla. Nuestra única protección contra las burbujas letales es la energía oscura, que evita que las más distantes lleguen hasta nosotros. De manera que, si abundan las burbujas letales, la energía oscura no será nuestra enemiga sino nuestra amiga.

¿Estamos solos?

Mucha gente da por descontado que existe vida avanzada en gran parte del universo, por lo que la extinción de la humanidad no tendría mucha importancia desde un punto de vista cósmico. Al fin y al cabo, ¿por qué habría de preocuparnos la posibilidad de que provocásemos nuestra propia extinción si una imponente civilización propia de *Star Trek* enseguida vendría y repoblaría nuestro sistema solar, quizá usando su tecnología avanzada para reconstruirnos y resucitarnos? Suponer que esto vaya a pasar me parece peligroso, porque puede generar en nosotros una falsa sensación de seguridad y hacer que nuestra civilización se vuelta apática e imprudente. De hecho, creo que la suposición de que no estamos solos en nuestro universo no solo es peligrosa, sino probablemente falsa.

Esta es una opinión minoritaria,(27) y puede que sea errónea, pero es al menos una posibilidad que hoy en día no podemos descartar, lo que nos impone el imperativo moral de ser prudentes y no conducir a nuestra civilización a la extinción.

Cuando doy conferencias sobre cosmología, suelo pedir a los asistentes que levanten la mano si creen que existe vida inteligente en algún otro lugar

de nuestro universo (la región del espacio desde la cual ha llegado hasta nosotros luz en los 13.800 millones de años transcurridos desde el Big Bang). Casi sin excepción, todo el mundo, desde alumnos de escuela infantil hasta estudiantes universitarios, alza la mano. Cuando pregunto por qué lo creen, la respuesta básica que obtengo es que nuestro universo es tan inmenso que tiene que haber vida en algún sitio, al menos desde un punto de vista estadístico. Analicemos con detenimiento este argumento y señalemos sus puntos flacos.

Todo se reduce a un solo número: la distancia típica entre una de las civilizaciones de la figura 6.10 y su vecina más cercana. Si esta distancia es mucho mayor de 20.000 millones de años luz, deberíamos esperar estar solos en nuestro universo (la parte del espacio desde la cual ha llegado hasta nosotros luz durante los 13.800 millones de años desde el Big Bang) y que nunca establezcamos contacto con alienígenas. ¿Qué valor cabe esperar que tenga esta distancia? No tenemos mucha idea. Esto significa que la distancia hasta nuestros vecinos es del orden de 1.000 … 000 metros, donde el número total de ceros podría ser 21, 22, 23, …, 100, 101, 102 o más, pero probablemente no mucho menor de 21, puesto que aún no hemos observado evidencia concluyente de presencia extraterrestre (véase la figura 6.11). Para que la civilización más próxima a nosotros esté dentro de nuestro universo, cuyo radio es del orden de 10^{26} metros, el número de ceros no puede ser mayor de 26, y la probabilidad de que dicha cifra esté en el estrecho margen entre 22 y 26 es bastante pequeña. Por eso creo que estamos solos en nuestro universo.

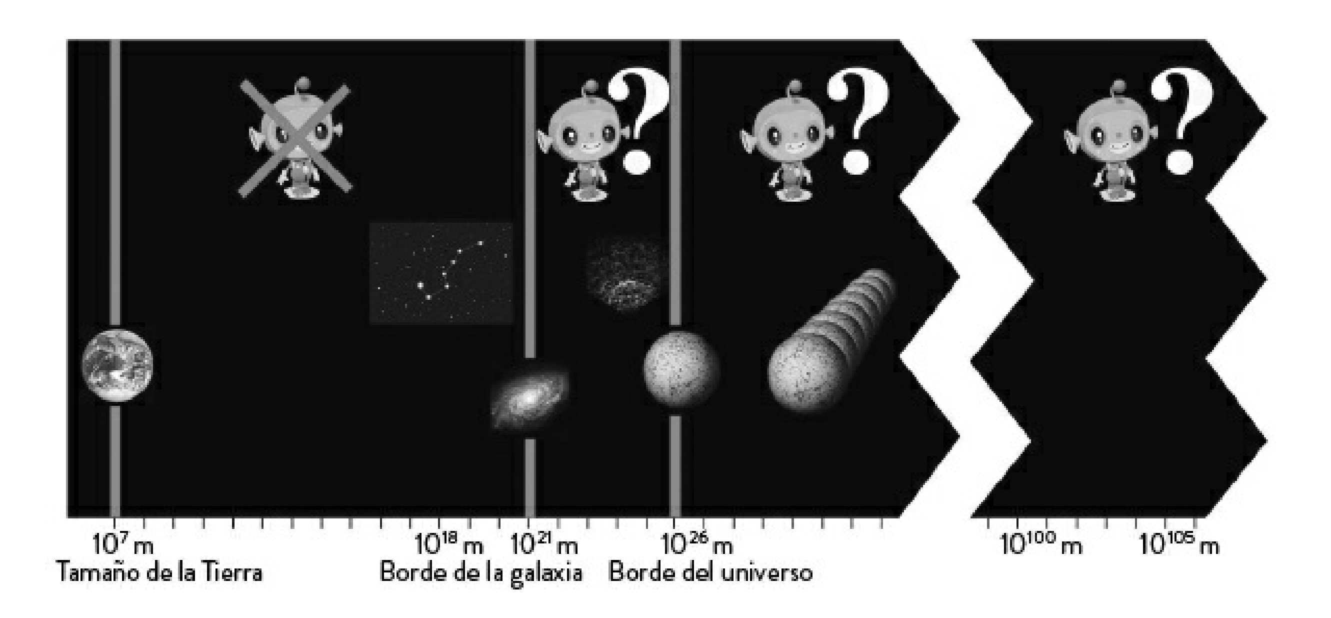


FIGURA 6.11. ¿Estamos solos? Las enormes incertidumbres sobre cómo evolucionaron la vida y la inteligencia sugieren que la civilización más cercana en el espacio podría estar razonablemente en cualquier punto del eje horizontal, lo que hace improbable que esté en el estrecho margen entre nuestra galaxia (a unos 10^{21} metros) y el borde de nuestro universo (a unos 10^{26} metros de distancia). Si estuviera mucho más cerca que este rango, debería haber tantas otras civilizaciones avanzadas en nuestra galaxia que probablemente las habríamos detectado, lo que sugiere que de hecho estamos solos en nuestro universo.

En mi libro *Nuestro universo matemático* ofrezco una justificación detallada de este argumento, por lo que no me extenderé aquí, pero la razón fundamental por la que desconocemos a qué distancia están nuestros vecinos es que, a su vez, no tenemos ni idea de cuál es la probabilidad de que surja vida inteligente en un determinado lugar. Como señaló el astrónomo estadounidense Frank Drake, esta probabilidad se puede calcular multiplicando la probabilidad de que en el lugar exista un entorno habitable (por ejemplo, un planeta con las características adecuadas), la probabilidad de que la vida surja allí y la probabilidad de que la vida evolucione hasta llegar a ser inteligente. Cuando yo estudiaba el doctorado, no teníamos ni idea de cómo calcular ninguna de estas probabilidades. Tras los espectaculares descubrimientos de las dos últimas décadas de planetas que orbitan alrededor de otras estrellas, ahora parece probable que abunden los planetas habitables, y que haya miles de millones solo en nuestra propia galaxia. Sin embargo, seguimos sin tener ni idea de cuál es la probabilidad de que surja la vida, y de que esta desarrolle inteligencia: algunos expertos creen que una o ambas son prácticamente inevitables y que ocurren en la mayoría de los planetas

habitables, mientras que otros piensan que una o ambas son extremadamente raras, debido a la existencia de uno o varios cuellos de botella evolutivos que requieren de un improbabilísimo golpe de suerte para producirse. Algunos de los cuellos de botella que se barajan implican problemas del tipo del huevo o la gallina en las primeras etapas de la vida capaz de reproducirse: por ejemplo, para que una célula moderna fabrique un ribosoma, la compleja maquinaria molecular que lee nuestro código genético y produce nuestras proteínas, necesita otro ribosoma, y no es nada evidente cómo pudo el primero de todos los ribosomas evolucionar gradualmente a partir de algo más simple.[89] Otros cuellos de botella identificados se refieren al desarrollo de una inteligencia superior. Por ejemplo, aunque los dinosaurios dominaron la Tierra durante más de cien millones de años, mil veces más tiempo del que nosotros, los humanos modernos, hemos existido, no parece que la evolución los llevase necesariamente a desarrollar una inteligencia superior y a inventar telescopios o computadoras.

Hay quien replica a mi argumento diciendo que es cierto que la vida inteligente podría ser algo excepcional, pero que de hecho no lo es: nuestra galaxia está repleta de vida inteligente de la que los científicos convencionales no se percatan. Quizá los extraterrestres ya hayan visitado nuestro planeta, como afirman los entusiastas de los ovnis. Quizá no hayan visitado la Tierra, pero están ahí fuera y se esconden de nosotros deliberadamente (es lo que el astrónomo estadounidense John A. Ball llama la «hipótesis del zoo», que aparece en clásicos de la ciencia ficción como *Hacedor de estrellas* de Olaf Stapledon). O quizá están ahí fuera sin tener la intención de ocultarse: simplemente no les interesa la colonización espacial o los grandes proyectos de ingeniería que hubiésemos sabido detectar.

Debemos mantener la mente abierta en torno a estas posibilidades, qué duda cabe, pero, puesto que no hay evidencia de ninguna de ellas que esté aceptada por la mayoría, también debemos considerar seriamente la alternativa: que estemos solos. Además, creo que no deberíamos subestimar la diversidad de las civilizaciones avanzadas y dar por supuesto que todas comparten objetivos que hacen que no las detectemos: vimos antes que es de lo más natural que una civilización tenga como objetivo la obtención de recursos, y para que lo detectásemos bastaría con que solo una civilización colonizase abiertamente el cosmos, incluida nuestra galaxia. Ante la evidencia de que en nuestra galaxia existen millones de planetas habitables

similares a la Tierra que son miles de millones de años más antiguos que nuestro planeta, con lo que sus ambiciosos habitantes habrían tenido tiempo más que suficiente para colonizar la galaxia, no podemos descartar la interpretación más evidente: que el origen de la vida requiere un evento aleatorio tan improbable que todos ellos están deshabitados.

Si, pese a lo anterior, la vida no es algo raro, pronto podríamos saberlo. Ambiciosos estudios astronómicos están analizando las atmósferas de los exoplanetas buscando evidencia de oxígeno de origen biológico. En paralelo con esta búsqueda de cualquier tipo de vida, la búsqueda de vida inteligente recibió recientemente un renovado impulso gracias a los cien millones de dólares que el filántropo ruso Yuri Milner ha donado al proyecto «Breakthrough Listen».

Es importante no caer en un excesivo antropocentrismo en la búsqueda de vida avanzada: si descubrimos una civilización extraterrestre, es probable que ya haya alcanzado la superinteligencia. Como explica Martin Rees en un ensayo reciente: «La historia de la civilización tecnológica humana se mide en siglos, y es posible que en solo uno o dos siglos más los humanos sean superados o trascendidos por la inteligencia inorgánica, que a partir de entonces persistirá y seguirá evolucionando durante miles de millones de años […] Sería hartamente improbable que la "pillásemos" en el breve lapso de tiempo en que adoptó forma orgánica».[90] Coincido con la conclusión a la que llega Jay Olson en su ya mencionado artículo científico sobre la colonización espacial: «Consideramos que la posibilidad de que una inteligencia avanzada haga uso de los recursos del universo simplemente para poblar planetas similares a la Tierra con versiones avanzadas de los humanos es un resultado final improbable de la progresión de la tecnología». Así pues, cuando pensemos en alienígenas, en lugar de imaginar hombrecitos verdes con dos brazos y dos piernas, pensemos mejor en la vida superinteligente que surca el espacio que vimos al principio del capítulo.

Aunque soy un firme defensor de los intentos actuales de encontrar vida extraterrestre, que están arrojando luz sobre una de las cuestiones más fascinantes de la ciencia, en privado espero que todos ellos fracasen y que no encuentren nada. La aparente incompatibilidad entre la abundancia de planetas habitables en nuestra galaxia y la ausencia de visitantes extraterrestres, conocida como *paradoja de Fermi*, sugiere la existencia de lo que el economista Robin Hanson llama un «gran filtro», un obstáculo

evolucionario/tecnológico en algún punto del recorrido evolutivo desde la materia inerte hasta la vida que coloniza el espacio. Si descubrimos que la vida surgió de forma independiente en algún otro lugar, esto sugeriría que la vida primitiva no es tan rara, y que el obstáculo se encuentra en algún momento posterior a nuestra fase actual de desarrollo (quizá porque la colonización espacial es imposible, o porque casi todas las civilizaciones avanzadas se autodestruyen antes de que puedan echarse al cosmos). Así pues, cruzo los dedos para que las búsquedas de vida extraterrestre no encuentren nada: esto sería compatible con el escenario en el que el desarrollo de vida inteligente es un evento raro, pero los humanos fuimos tan afortunados que hemos dejado atrás ese obstáculo y tenemos ante nosotros un extraordinario potencial.

PERSPECTIVA

Hasta ahora, en este libro nos hemos dedicado a repasar la historia de la vida en el universo, desde sus humildes orígenes hace miles de millones de años hasta posibles futuros gloriosos dentro de otros miles de millones de años más. Si el desarrollo actual de la IA acaba desencadenando una explosión de inteligencia y la colonización optimizada del espacio, será una explosión en un sentido verdaderamente cósmico: tras pasar miles de millones de años como una perturbación casi despreciable en un cosmos inerte e indiferente, la vida explota de forma súbita en el escenario cósmico como una onda expansiva esférica que se propaga casi a la velocidad de la luz, nunca se detiene, y prende allá por donde pasa la llama de la vida.

Muchos de los pensadores que han ido apareciendo a lo largo del libro han articulado de forma elocuente visiones igualmente optimistas sobre la importancia de la vida en nuestro futuro cósmico. Debido a que a menudo se desdeña a los autores de ciencia ficción como soñadores románticos y poco realistas, encuentro paradójico que la mayoría de los escritos tanto de ciencia ficción como científicos sobre la colonización espacial parezcan ahora demasiado *pesimistas* a la luz de la superinteligencia. Por ejemplo, vimos cómo los viajes intergalácticos serán mucho más fáciles una vez que las personas y otras entidades inteligentes puedan transmitirse en forma digital, lo que nos convertiría en dueños de nuestro propio destino no solo en el

sistema solar o en la Vía Láctea, sino también en el cosmos.

Anteriormente consideramos la posibilidad muy real de que seamos la única civilización con tecnología avanzada en todo nuestro universo. Dedicaremos el resto de este capítulo a explorar este escenario y la enorme responsabilidad moral que conlleva. Esto significa que, tras 13.800 millones de años, la vida en nuestro universo ha llegado a una bifurcación en el camino, y debe elegir entre florecer en todo el cosmos o extinguirse. Si no seguimos mejorando nuestra tecnología, la cuestión no es si la humanidad se extinguirá, sino cómo. ¿Qué acabará antes con nosotros: un asteroide, un supervolcán, el calor abrasador del Sol en las postrimerías de su vida o alguna otra calamidad (véase la figura 5.1)? Una vez que hayamos desaparecido, el drama cósmico predicho por Freeman Dyson proseguirá sin espectadores: salvo que ocurra un cosmocalipsis, las estrellas se apagarán, las galaxias se desvanecerán y los agujeros negros se evaporarán, despidiéndose cada uno de ellos con una gran explosión que liberará más de un millón de veces la energía de la Bomba del Zar, la bomba de hidrógeno más potente jamás construida. Como dijo Freeman: «Un universo frío y en expansión será iluminado por fuegos artificiales ocasionales durante mucho tiempo». Por desgracia, este espectáculo de fuegos artificiales será un derroche sin sentido, ya que no habrá nadie para disfrutarlo.

Sin tecnología, la extinción humana es inminente en el contexto cósmico de decenas de miles de millones de años, lo que haría que todo el drama de la vida en el universo quedase meramente en un breve y pasajero destello de belleza, pasión y significado en una cuasi eternidad sin ningún sentido que nadie experimentará. ¡Qué oportunidad perdida sería esta! Si en vez de rehuir la tecnología, optamos por sacarle todo el provecho, lo que hacemos es subir la apuesta y abrir la posibilidad tanto de que la vida sobreviva y florezca, como de que se extinga incluso antes, autodestruyéndose debido a una deficiente planificación (véase la figura 5.1). Yo me inclino por apostar por la tecnología y proceder no con fe ciega en lo que construimos, sino con cautela, previsión y una cuidadosa planificación.

Después de 13.800 millones de años de historia cósmica, estamos ante un universo de una belleza sobrecogedora, que a través de nosotros los humanos ha cobrado vida y ha comenzado a tomar conciencia de sí mismo. Hemos visto que el potencial futuro de la vida en el universo es más halagüeño que los sueños más febriles de nuestros antepasados, aunque atemperado por la

posibilidad igualmente real de que la vida inteligente se extinga permanentemente. ¿Desarrollará la vida en el universo todo su potencial o lo desperdiciará? Esto depende en gran medida de lo que hagamos durante nuestras vidas los humanos que estamos vivos actualmente, y tengo confianza en que podamos hacer que el futuro de la vida sea realmente maravilloso si tomamos las decisiones correctas. ¿Qué deberíamos querer y cómo podemos lograr esos objetivos? Dedicaremos el resto del libro a analizar algunas de las dificultades más complicadas para alcanzarlos, y qué podemos hacer al respecto.

CONCLUSIONES

- En comparación con las escalas de tiempo cósmicas de miles de millones de años, una explosión de inteligencia es un evento repentino donde la tecnología se estabiliza rápidamente en un nivel limitado solo por las leyes de la física.
- Esta meseta tecnológica es mucho más elevada que la tecnología actual, y hace posible que una cantidad dada de materia genere aproximadamente diez mil millones de veces más energía (usando esfalerones o agujeros negros), almacene una cantidad de información entre 12 y 18 órdenes de magnitud mayor o realice cálculos a una velocidad entre 31 y 41 órdenes de magnitud más rápida, o que pueda convertirse a cualquier otra forma de materia que se desee.
- La vida superinteligente no solo haría un uso mucho más eficiente de los recursos existentes, sino que podría hacer crecer la biosfera actual en aproximadamente 32 órdenes de magnitud al obtener más recursos a través de la colonización cósmica a una velocidad cercana a la de la luz.
- La energía oscura limita la expansión cósmica de la vida superinteligente y también la protege de remotas burbujas letales en expansión o de las civilizaciones hostiles. La amenaza de que la energía oscura desgarre las civilizaciones cósmicas motiva gigantescos proyectos de ingeniería cósmica, incluida la construcción de agujeros de gusano si esto resulta ser factible.
- El principal producto que se compartirá o con el que se comerciará a través de distancias cósmicas probablemente será la información.
- Exceptuando el caso de los agujeros de gusano, la existencia de un límite de velocidad en la comunicación, dado por la velocidad de la luz, plantea serios desafíos para la coordinación y el control en una civilización cósmica. Un centro de operaciones distante puede incentivar a sus «nodos» superinteligentes para que cooperen mediante el uso de recompensas o de amenazas, por ejemplo mediante el despliegue de una IA guardiana local programada para destruir el nodo activando una supernova o un cuásar a menos que se obedezcan las reglas.
- La colisión de dos civilizaciones en expansión puede dar como resultado la asimilación, la cooperación o la guerra. Esta última posibilidad será menos probable que entre las civilizaciones actuales.
- A pesar de la creencia popular en sentido contrario, es bastante posible que seamos la única forma de vida capaz de hacer que nuestro universo observable cobre vida en el futuro.
- Si no mejoramos nuestra tecnología, la cuestión no es si la humanidad se extinguirá, sino simplemente cómo: ¿acabará antes con nosotros un asteroide, un supervolcán, el calor abrasador

de un Sol en las postrimerías de su vida, o alguna otra calamidad?

- Si seguimos mejorando nuestra tecnología con suficiente cuidado, previsión y capacidad de planificación para evitar los riesgos, la vida tiene el potencial de florecer en la Tierra y mucho más allá durante muchos miles de millones de años, superando incluso los sueños más descabellados de nuestros antepasados.

7
OBJETIVOS

El secreto de la existencia humana no solo está en vivir, sino también en saber para qué se vive.

FIÓDOR DOSTOIEVSKI, *Los hermanos Karamázov*

La vida es un viaje, no un destino.

RALPH WALDO EMERSON

Si tuviese que resumir en una sola palabra sobre qué giran las polémicas más espinosas relacionadas con la IA, esta palabra sería «objetivos»: ¿debemos darle objetivos a la IA, y, de ser así, los objetivos de quién? ¿Cómo podemos darle objetivos? ¿Podemos asegurarnos de que la IA conserva esos objetivos a medida que aumenta su inteligencia? ¿Podemos modificar los objetivos de una IA que sea más inteligente que nosotros? ¿Cuáles son nuestros objetivos últimos? Estas preguntas no solo son difíciles, sino también cruciales para el futuro de la vida: si no sabemos lo que queremos, es menos probable que lo consigamos, y, si cedemos el control a máquinas que no comparten nuestros objetivos, es más probable que lo que obtengamos sea algo que no queremos.

FÍSICA: EL ORIGEN DE LOS OBJETIVOS

Para arrojar luz sobre estas cuestiones, primero analizaremos cuál es, en última instancia, el origen de los objetivos. Cuando observamos el mundo que nos rodea, nos parece que algunos procesos son *intencionales*, mientras que otros no. Consideremos, por ejemplo, el golpe a un balón de fútbol para marcar el gol que decide el partido. El comportamiento del balón en sí no parece que sea intencional, y la manera más económica de explicarlo es a través de las leyes del movimiento de Newton, como reacción a la patada que recibe. Por otra parte, la forma más económica de explicar el comportamiento

de la jugadora no es de tipo mecanicista, en función de átomos que se empujan los unos a los otros, sino por el *propósito* que ella tiene de maximizar el tanteo de su equipo. ¿Cómo surgió ese comportamiento intencional a partir de la física del universo primitivo, que consistía únicamente en un montón de partículas que rebotaban de un lado a otro sin ningún propósito aparente?

Curiosamente, las raíces últimas del comportamiento intencional pueden rastrearse en las propias leyes físicas, y se manifiestan incluso en procesos simples que nada tienen que ver con la vida. Si una socorrista rescata a un bañista, como en la figura 7.1, no esperamos que vaya hacia él en línea recta, sino que corra un poco más a lo largo de la orilla, por donde puede ir más rápido que por el agua, y se desvíe ligeramente al lanzarse al mar. Interpretamos de forma natural que la trayectoria que sigue es intencional, ya que, de todas las posibles, está eligiendo deliberadamente la óptima para llegar hasta el bañista en el menor tiempo posible. Ahora bien, un simple rayo de luz se desvía de manera similar cuando entra en el agua (véase la figura 7.1), minimizando así también el tiempo que tarda en llegar a su destino. ¿Cómo es posible?

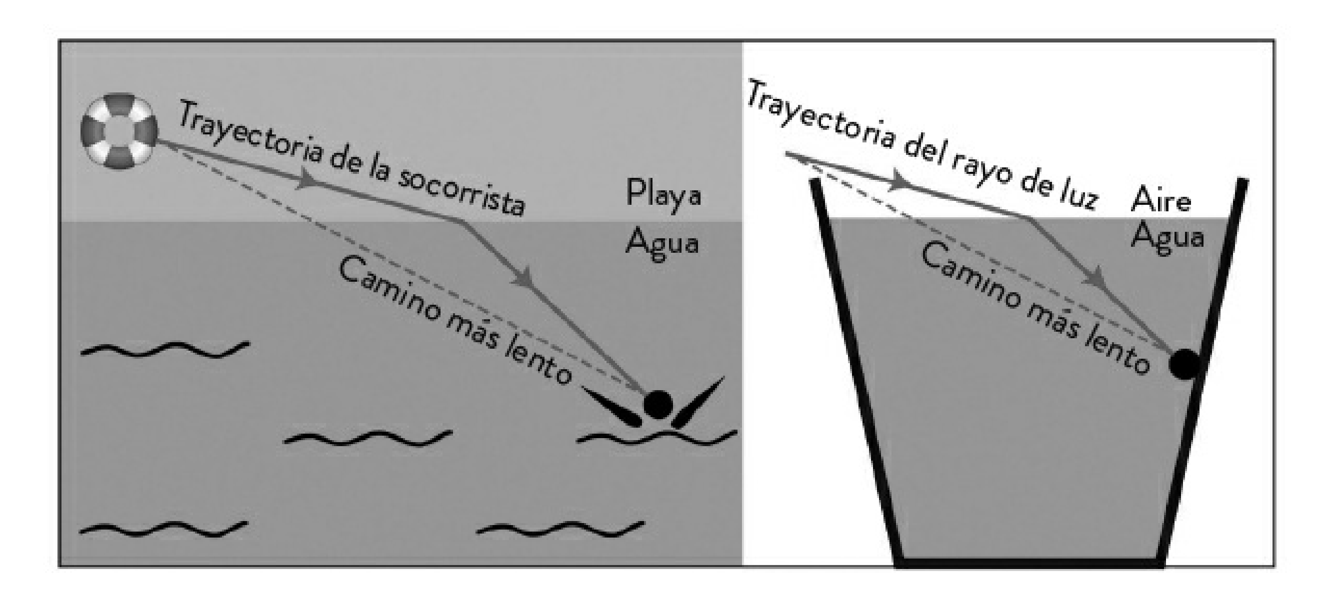


FIGURA 7.1. Para rescatar a un bañista lo más rápido posible, una socorrista no irá en línea recta (línea de puntos), sino que hará más parte del recorrido por la playa, por donde puede ir más rápido que por el agua. Análogamente, un rayo de luz se desvía al entrar en el agua para llegar antes a su destino.

Se trata de lo que se conoce como *principio de Fermat*, formulado en

1662, y ofrece una forma alternativa de predecir el comportamiento de los rayos de luz. Es notable cómo los físicos han ido descubriendo desde entonces que todas las leyes de la física clásica pueden reformularse matemáticamente de una manera análoga: de todas las formas en que la naturaleza podría elegir hacer algo, prefiere la óptima, lo que por lo general se reduce a minimizar o maximizar alguna magnitud. Existen dos maneras matemáticamente equivalentes de describir cada ley física: expresando que el pasado causa el presente, o que la naturaleza optimiza alguna cosa. Aunque la segunda manera no se suele explicar en los cursos introductorios de física, porque las matemáticas son más difíciles, creo que es más elegante y profunda. Si una persona está intentando optimizar algo (por ejemplo, su tanteo, su riqueza o su felicidad) describiremos naturalmente su empeño como intencional. Por lo que, si es la propia naturaleza la que intenta optimizar algo, no es de extrañar que pueda manifestarse un comportamiento intencional: estaba latente desde el principio, en las mismísimas leyes de la física.

Una famosa magnitud que la naturaleza se empeña en maximizar es la *entropía*, que en términos generales mide el grado de desorden de las cosas. La segunda ley de la termodinámica afirma que la entropía tiende a aumentar hasta que alcanza el máximo valor posible. Si ignoramos por el momento los efectos de la gravedad, ese estado de desorden máximo se conoce como *muerte térmica*, y corresponde a una situación en la que todo está distribuido en una aburrida uniformidad perfecta, sin complejidad, sin vida y sin cambios. Por ejemplo, cuando vertemos leche fría en un café caliente, nuestra bebida tiende irreversiblemente a su propia forma de muerte térmica, y al cabo de poco tiempo toda ella es una mezcla uniforme y tibia. Si un organismo vivo muere, su entropía también comienza a crecer, y enseguida la disposición de sus partículas se va volviendo mucho menos ordenada.

El propósito evidente de la naturaleza de incrementar la entropía ayuda a explicar por qué parece que el tiempo tiene un sentido preferido, lo que hace que las películas no resulten realistas cuando se reproducen hacia atrás: si se nos cae una copa de vino, esperamos que se haga añicos contra el suelo y que el desorden total (la entropía) aumente. Si a continuación viésemos cómo esos añicos *se recomponen* y la copa vuelve volando hasta nuestra mano (con la consiguiente disminución de la entropía), probablemente no nos la tomaríamos, y pensaríamos que ya hemos bebido más de la cuenta.

Cuando me di cuenta por primera vez de nuestra inexorable marcha hacia la muerte térmica, me pareció algo bastante deprimente; y no solo me ocurría a mí: el pionero de la termodinámica lord Kelvin escribió en 1841 que «el resultado iba a ser inevitablemente un estado de reposo y muerte universales». Cuesta encontrar consuelo en la idea de que el objetivo a largo plazo de la naturaleza consiste en maximizar la muerte y la destrucción. Sin embargo, descubrimientos más recientes demuestran que las cosas no pintan tan mal. Para empezar, la gravedad se comporta de manera distinta a todas las demás fuerzas, y se esfuerza por hacer que el universo no sea más uniforme y aburrido, sino más heterogéneo e interesante. Así pues, la gravedad transformó el aburrido universo primitivo, casi del todo uniforme, en el cosmos actual, heterogéneo y maravillosamente complejo, repleto de galaxias, estrellas y planetas. Gracias a la gravedad, hay ahora una gran variedad de temperaturas, lo que hace posible que la vida florezca al combinar calor y frío: vivimos en un planeta agradablemente templado que absorbe calor solar a 6.000 °C y se enfría al radiar calor residual al gélido espacio, cuya temperatura es de apenas 3 °C por encima del cero absoluto.

En segundo lugar, el trabajo reciente de mi colega en el MIT Jeremy England junto a otros investigadores ha traído más noticias buenas, al demostrar que la termodinámica también dota a la naturaleza de un propósito más estimulante que la muerte térmica.[91] Este propósito recibe el nombre técnico de *adaptación motivada por la disipación*, lo que significa básicamente que grupos aleatorios de partículas tratan de organizarse entre sí para extraer energía de su entorno de la manera más eficiente posible («disipación» es hacer que aumente la entropía, por lo general al transformar energía útil en calor, a menudo mediante la realización de trabajo útil). Por ejemplo, un montón de moléculas expuestas a la luz del sol con el tiempo tenderían a reorganizarse para absorber dicha luz cada vez con mayor eficiencia. Dicho de otro modo, la naturaleza parece que lleva incorporado el propósito de producir sistemas con capacidad de autoorganización cada vez más complejos y más cercanos a la vida, y este propósito está integrado en todas las leyes de la física.

¿Cómo podemos reconciliar este impulso cósmico hacia la vida con el impulso cósmico hacia la muerte? Podemos encontrar la respuesta en el famoso libro *¿Qué es la vida?*, publicado en 1944 por Erwin Schrödinger, uno de los fundadores de la mecánica cuántica. En él, Schrödinger señala que

una de las características distintivas de un sistema vivo es que mantiene o reduce su entropía haciendo que aumente la de su entorno. En otras palabras, la segunda ley de la termodinámica incluye una salvedad para la vida: aunque la entropía total debe aumentar, está permitido que dicha magnitud disminuya en ciertos lugares, siempre que crezca en mayor medida en otros. Así pues, la vida mantiene o incrementa su complejidad al hacer que aumente el desorden en su entorno.

BIOLOGÍA: LA EVOLUCIÓN DE LOS OBJETIVOS

Acabamos de ver cómo el origen del comportamiento intencional puede rastrearse hasta las leyes físicas, que parecen dotar a las partículas con el propósito de organizarse de manera que extraigan energía de su entorno con la mayor eficiencia posible. Una excelente forma en que la disposición de las partículas puede aproximarse a ese objetivo consiste en hacer copias de sí misma para producir más captadores de energía. Hay muchos ejemplos bien conocidos de esta emergencia de autorreproducción: por ejemplo, los vórtices en un fluido turbulento pueden crear copias de sí mismos, y los cúmulos de microesferas pueden alentar a otras microesferas cercanas a crear cúmulos idénticos. En un momento dado, alguna disposición particular de partículas es tan eficiente a la hora de generar copias de sí misma que podría hacerlo casi indefinidamente a base de extraer energía y materias primas de su entorno. Esa disposición de partículas es lo que llamamos *vida*. Aún sabemos muy poco sobre cómo se originó la vida en la Tierra, pero sí sabemos que ya existían formas de vida primitivas hace unos 4.000 millones de años.

Si una forma de vida crea copias de sí misma, y estas copias hacen lo propio, el número total irá duplicándose regularmente hasta que la población alcance un tamaño tal que ponga de manifiesto las limitaciones de recursos y otros problemas. La reproducción sostenida produce enseguida cantidades inmensas: si empezamos con uno y lo multiplicamos por dos sucesivamente solo trescientas veces, obtenemos una cantidad superior al número de partículas existentes en el universo. Esto significa que, no mucho tiempo después de que apareciese la primera forma de vida primitiva, cantidades enormes de materia habían cobrado vida. Algunas de las copias no eran perfectas, por lo que enseguida hubo muchas formas de vida diferentes

intentando replicarse, compitiendo entre sí por los mismos recursos finitos. La evolución darwiniana había comenzado.

Si hubiésemos estado observando discretamente la Tierra en la época en que surgió la vida, habríamos notado un cambio drástico en el comportamiento intencional. Mientras que antes parecía que las partículas intentaban incrementar el desorden promedio de distintas maneras, los patrones autorreplicantes que acababan de volverse ubicuos parecían tener un propósito diferente: no la disipación sino la *reproducción*. Charles Darwin explicó con elegancia el porqué: puesto que los replicantes más eficientes superaban y se imponían a los demás, enseguida cualquier forma de vida aleatoria en la que nos fijásemos estaría altamente optimizada para el propósito de la reproducción.

¿Cómo pudo el objetivo pasar de la disipación a la reproducción cuando las leyes físicas son las mismas? La respuesta es que el objetivo fundamental (disipación) no cambió, sino que condujo a un *objetivo instrumental* distinto, esto es, un subobjetivo que ayudó a lograr el objetivo fundamental. Pensemos, por ejemplo, en la alimentación. Aparentemente, todos tenemos el propósito de aplacar la sensación de hambre, aunque sabemos que el único propósito fundamental de la evolución es la reproducción, no el masticar. Esto se debe a que comer contribuye a la reproducción: morirse de hambre impide tener hijos. De manera similar, la reproducción contribuye a la disipación, porque un planeta repleto de vida es más eficiente a la hora de disipar energía. Así pues, en cierto sentido, el cosmos inventó la vida como medio para acelerar su evolución hacia la muerte térmica. Si tiramos azúcar por el suelo de la cocina, en principio podría conservar durante años su energía química útil, pero, si aparecen hormigas, disiparán esa energía en un abrir y cerrar de ojos. Análogamente, las reservas de petróleo enterradas en la corteza terrestre habrían conservado su energía química útil durante mucho más tiempo si las formas de vida bípedas no las hubiésemos extraído y las hubiésemos quemado.

Entre los habitantes actuales de la Tierra, productos de la evolución, estos propósitos instrumentales parecen haber cobrado vida propia: aunque la evolución los optimizó para el único propósito de la reproducción, muchos dedican buena parte de su tiempo no a generar descendencia, sino a actividades como dormir, buscar comida, construir hogares, ejercer su dominio y combatir o ayudar a los demás (en ocasiones hasta extremos que

llegan a *reducir* la población). La investigación en psicología evolutiva, economía e inteligencia artificial ha explicado con elegancia por qué. Algunos economistas solían representar a las personas como agentes racionales, sujetos decisores idealizados que siempre elegían la acción que fuese óptima para la consecución de su propósito, pero, evidentemente, esto no es realista. En la práctica, estos agentes tienen lo que Herbert Simon, pionero de la IA y ganador del Premio Nobel, denominaba «racionalidad limitada», porque sus recursos son también limitados: la racionalidad de sus decisiones está limitada por la información con la que cuentan, el tiempo del que disponen y el hardware que tienen para pensar. Esto significa que, cuando la evolución darwiniana está optimizando un organismo para que alcance un propósito, lo máximo que puede hacer es implementar un algoritmo aproximado que funcione razonablemente bien en el contexto restringido en el que el agente se encuentra de forma habitual. La evolución ha implementado la optimización para la reproducción de esta manera: en lugar de plantearse en cada situación cuál será la acción que maximizará el número de descendientes viables de un organismo, implementa una mezcolanza de apaños heurísticos: reglas empíricas que normalmente suelen funcionar. Para la mayoría de los animales son, entre otras, el impulso sexual, la necesidad de beber cuando están sedientos y de comer cuando están hambrientos, así como la tendencia a evitar las cosas que tienen mal sabor o hacen daño.

Estas reglas empíricas fracasan a veces estrepitosamente en situaciones para las que no estaban diseñadas, como cuando una rata come un veneno delicioso, cuando las polillas se sienten atraídas hacia trampas pegajosas por fragancias seductoras propias de las hembras o los insectos vuelan hacia la llama de una vela hasta abrasarse en ella.(28) Puesto que la sociedad humana actual es muy distinta del entorno para el cual la evolución optimizó nuestras reglas empíricas, no debería sorprendernos saber que en ocasiones nuestro comportamiento no tiene como resultado maximizar la procreación. Por ejemplo, el subobjetivo de no morir de hambre se implementa en parte como un deseo de consumir alimentos calóricos, lo que ha dado lugar a la actual epidemia de obesidad y a las dificultades para encontrar pareja. El subobjetivo de procrear se implementó como un deseo sexual, en lugar de como un deseo de ser donante de esperma/óvulos, aunque este último puede producir más bebés con menos esfuerzo.

En resumen, un organismo vivo es un agente de racionalidad limitada que no persigue un único propósito, sino que sigue reglas empíricas para decidir qué perseguir y qué evitar. Nuestra mente humana percibe estas reglas empíricas fruto de la evolución como *sentimientos*, que normalmente guían nuestra toma de decisiones hacia el propósito final de la reproducción (muchas veces incluso sin que seamos conscientes de ello). La sensación de hambre y de sed nos protege de la inanición y la deshidratación, la sensación de dolor nos protege de dañar nuestros cuerpos, la sensación de lujuria nos impulsa a procrear, los sentimientos de amor y compasión nos llevan a ayudar a otros individuos portadores de nuestros genes y a quienes los ayudan, etcétera. Guiados por estos sentimientos, nuestros cerebros pueden decidir rápida y eficientemente qué hacer sin tener que analizar en detalle todas las opciones en función de su repercusión sobre el número de descendientes que produciremos. Para perspectivas estrechamente relacionadas con los sentimientos y sus raíces fisiológicas, recomiendo encarecidamente los escritos de William James y António Damásio.[92]

Es importante señalar que cuando, en ocasiones, nuestros sentimientos operan contra la procreación, no se trata necesariamente de algo accidental o de que estemos engañándonos: nuestro cerebro puede rebelarse contra nuestros genes y su propósito de reproducción de forma deliberada, por ejemplo al elegir usar métodos anticonceptivos. Otros ejemplos más extremos de la rebelión del cerebro contra sus genes incluyen la decisión de suicidarse o de pasar la vida célibe para ser cura, monje o monja.

¿Por qué optamos en ocasiones por rebelarnos contra nuestros genes y su propósito de reproducirse? Lo hacemos porque, por cómo estamos diseñados, como agentes de racionalidad limitada, solo somos fieles a nuestros sentimientos. Aunque nuestro cerebro evolucionó simplemente para ayudar a que nuestros genes se replicasen, no le importa en absoluto este propósito, ya que no tenemos sentimientos relacionados con nuestros genes (de hecho, durante casi toda la historia de la humanidad, nuestros antepasados ni siquiera sabían que tenían genes). Además, el cerebro es mucho más inteligente que los genes y, ahora que comprendemos cuál es el propósito de

estos (la reproducción), nos parece bastante banal y esquivable. Las personas pueden darse cuenta de que los genes les hacen sentir lujuria, pero al mismo tiempo tener muy pocas ganas de criar a quince hijos, de manera que hackean su programación genética al combinar la gratificación emocional de las relaciones íntimas con los métodos anticonceptivos. Puede que sean conscientes de que los genes estimulan el deseo de algo dulce, pero que al mismo no quieran ganar peso, y por tanto opten por hackear su programación genética al combinar la gratificación emocional de una bebida azucarada con las cero calorías de los edulcorantes artificiales.

Aunque el hackeo de los mecanismos de gratificación en ocasiones sale mal, como cuando la gente se hace adicta a la heroína, nuestro patrimonio genético ha conseguido sobrevivir hasta ahora sin problemas, a pesar de nuestro cerebro astuto y rebelde. Conviene recordar, no obstante, que la autoridad última recae ahora en nuestros sentimientos, no en nuestro cerebro. Esto significa que el comportamiento humano no está estrictamente optimizado para la supervivencia de nuestra especie. De hecho, puesto que nuestros sentimientos se limitan a implementar reglas empíricas que no son apropiadas para todas las situaciones, el comportamiento humano, en sentido estricto, no tiene en absoluto un propósito único y bien definido.

INGENIERÍA: EXTERNALIZAR LOS OBJETIVOS

¿Pueden las máquinas tener propósitos? Esta sencilla pregunta ha desatado grandes polémicas, porque su significado se presta a distintas interpretaciones, a menudo relacionadas con asuntos espinosos como si las máquinas pueden o no ser conscientes o tener sentimientos. Pero si somos más prácticos y simplemente entendemos que lo que la pregunta significa es «¿Pueden las máquinas exhibir comportamiento intencional?», la respuesta es obvia: «Por supuesto que pueden, puesto que podemos diseñarlas así». Diseñamos las ratoneras para que tengan el propósito de atrapar ratones, lavavajillas con el propósito de limpiar los platos, y relojes con el propósito de marcar la hora. Cuando nos enfrentamos a una máquina, normalmente lo único que nos importa es el hecho empírico de que exhibe un comportamiento intencional: cuando nos persigue un misil que se guía por el calor, nos importa bien poco saber si tiene sentimientos o si es consciente. Si

le sigue resultando incómodo decir que el misil tiene un propósito aunque no tenga consciencia, de momento puede usted simplemente leer «propósito» donde dice «objetivo» y viceversa (en el capítulo siguiente abordaremos la cuestión de la consciencia).

Hasta ahora, la mayoría de las cosas que construimos exhibe únicamente un *diseño* —no un *comportamiento*— intencional: una autopista no se comporta de ninguna manera, simplemente está ahí. Sin embargo, la explicación más económica para su existencia es que fue diseñada para lograr un propósito, por lo que incluso una tecnología pasiva como esta está contribuyendo a que el universo sea más intencional. La *teleología* es la explicación de las cosas en función de sus propósitos, y no de sus causas, de manera que podemos resumir la primera parte de este capítulo diciendo que el universo es cada vez más teleológico.

La materia inerte no solo puede tener propósitos, al menos en este sentido débil, sino que cada vez los tiene en mayor medida. Si hubiésemos observado los átomos que componen la Tierra desde su formación, habríamos asistido a tres fases de comportamiento intencional:

1. Toda la materia parecía centrada en la disipación (aumento de la entropía).
2. Parte de la materia cobró vida y pasó a centrarse en la reproducción y en subobjetivos derivados de ella.
3. Una parte rápidamente creciente de la materia fue reorganizada por los seres vivos para ayudarlos a conseguir sus propósitos.

La tabla 7.1 muestra en qué medida la humanidad ha llegado a ser dominante desde un punto de vista físico: no solo sumamos más materia que todos los mamíferos excepto las vacas (que son tan numerosas porque sirven a nuestros propósitos de consumir carne y productos lácteos), sino que la materia que compone nuestras máquinas, carreteras, edificios y otros proyectos de ingeniería parece ir camino de superar a toda la materia viva existente en la Tierra. Dicho de otro modo, incluso sin una explosión de inteligencia, la mayor parte de la materia existente en la Tierra que exhibe propiedades intencionales pronto podría ser fruto del diseño, y no de la evolución.

ENTIDADES INTENCIONALES	MILES DE MILLONES DE TONELADAS
5×10^{30} bacterias	400

Plantas	400
10^{15} mesophelagic fish	10
$1,3 \times 10^9$ vacas	0,5
7×10^9 humanos	0,4
10^{14} hormigas	0,3
$1,7 \times 10^6$ ballenas	0,0005
Hormigón	100
Acero	20
Asfalto	15
$1,2 \times 10^8$ coches	2

TABLA 7.1. Cantidades aproximadas de materia en la Tierra en forma de entidades que han evolucionado o han sido diseñadas para tener un propósito. Las entidades diseñadas, como los edificios, las carreteras o los automóviles parecen encaminadas a superar a las entidades que son fruto de la evolución, como las plantas y los animales.

Esta nueva tercera clase de comportamiento intencional tiene el potencial de ser mucho más variada que las precedentes: mientras que las entidades producto de la evolución tienen todas el mismo propósito último (la reproducción), las entidades diseñadas pueden tener prácticamente cualquier propósito último, incluso propósitos opuestos. Los fogones tratan de calentar los alimentos, mientras que los frigoríficos procuran enfriarlos. Los generadores intentan transformar el movimiento en electricidad, mientras que los motores buscan convertir la electricidad en movimiento. Los programas de ajedrez normales intentan ganar al ajedrez, pero también existen otros que compiten en torneos con el propósito de perder al ajedrez.

Hay una tendencia histórica a que las entidades diseñadas asuman objetivos que no son solo más variados, sino también más *complejos*: nuestros dispositivos son cada vez más inteligentes. Diseñamos nuestras primeras máquinas y otros artefactos dotados de objetivos muy simples: por ejemplo, casas para mantenernos calientes, secos y seguros. Progresivamente, hemos aprendido a construir máquinas con objetivos más complejos, como aspiradores robotizados, o cohetes y coches autónomos. Los avances recientes en IA nos han proporcionado sistemas como Deep Blue, Watson y AlphaGo, cuyos objetivos de ganar al ajedrez, imponerse en los concursos televisivos o vencer al go son tan complicados que se requiere una

considerable maestría humana para apreciar debidamente lo competentes que son tales sistemas.

Cuando construimos una máquina para que nos ayude, puede resultar difícil hacer que sus propósitos se alineen perfectamente con los nuestros. Por ejemplo, una ratonera puede confundir los dedos de su pie con un roedor hambriento, con dolorosas consecuencias. Todas las máquinas son agentes con racionalidad limitada, e incluso las máquinas actuales más sofisticadas tienen una comprensión del mundo más pobre que la nuestra, por lo que las reglas, que utilizan para decidir lo que debe hacerse, a menudo son demasiado simplistas. Esa ratonera salta con demasiada facilidad porque no tiene ni idea de lo que es un ratón, muchos accidentes industriales mortales ocurren porque las máquinas no tienen ni idea de lo que es una persona, y los ordenadores que provocaron la «quiebra instantánea» de un billón de dólares en Wall Street en 2010 desconocían que lo que estaban haciendo no tenía ningún sentido. Por lo tanto, muchos de estos problemas de alineación de objetivos podrían resolverse haciendo que nuestras máquinas fuesen más inteligentes, pero, como hemos aprendido con Prometeo en el capítulo 4, el hecho de que las máquinas sean cada vez más inteligentes puede plantear nuevas e importantes dificultades para asegurarse de que comparten nuestros objetivos.

IA AMIGABLE: CONFORMAR OBJETIVOS

Cuanto más inteligentes y poderosas sean las máquinas, más importante será que sus objetivos estén alineados con los nuestros. Mientras solo construyamos máquinas relativamente tontas, la cuestión no será si los objetivos humanos acabarán prevaleciendo, sino solamente cuántos problemas podrían causar estas máquinas a la humanidad antes de que encontrásemos la manera de resolver el problema de alineación de los objetivos. Sin embargo, si alguna vez se libera una superinteligencia, las tornas se invertirán: puesto que la inteligencia es la capacidad de lograr objetivos, una IA superinteligente es, por definición, mucho más capaz de lograr sus objetivos que los humanos de alcanzar los nuestros, y por tanto sería aquella la que se impondría. En el capítulo 4 vimos muchos ejemplos de una situación así con Prometeo como protagonista. Si en este momento

queremos experimentar lo que sucede cuando los objetivos de una máquina se imponen a los nuestros, basta con que descarguemos un programa de ajedrez puntero y tratemos de vencerle. No lo conseguiremos, y enseguida nos aburriremos intentándolo…

En otras palabras, *el peligro real de la* IAG *no radica en su maldad sino en su competencia*. Una IA superinteligente será muy capaz de alcanzar sus objetivos, y si esos objetivos no están alineados con los nuestros tendremos problemas. Como mencioné en el capítulo 1, las personas no tienen ningún inconveniente en anegar hormigueros para construir centrales hidroeléctricas. Evitemos, pues, poner a la humanidad en el lugar de las hormigas. La mayoría de los investigadores argumentan que, si alguna vez acabamos creando una superinteligencia, debemos asegurarnos de que se trata de lo que el pionero de la IA segura Eliezer Yudkowsky ha bautizado como «IA amigable»: IA cuyos objetivos estén alineados con los nuestros.[93]

Encontrar la manera de conformar los objetivos de una IA superinteligente con los nuestros no solo es importante, sino también difícil. De hecho, es un problema que aún está por resolver. Se divide en tres subproblemas complicados, cada uno de los cuales es objeto de investigación activa por parte de informáticos y otros pensadores:

1. Hacer que la IA *entienda* nuestros objetivos
2. Hacer que la IA *adopte* nuestros objetivos
3. Hacer que la IA *conserve* nuestros objetivos

Analicémoslos uno por uno, dejando para la siguiente sección la cuestión de qué hemos de entender por «nuestros objetivos».

Para entender nuestros objetivos, una IA debe dilucidar no lo que hacemos sino por qué lo hacemos. A los humanos nos cuesta tan poco esfuerzo hacer esto que es fácil olvidar lo difícil que es esta tarea para un ordenador, y lo fácil que es de malinterpretar. Si le pidiésemos a un futuro coche autónomo que nos llevase al aeropuerto lo más rápidamente posible, y este interpretase nuestra solicitud literalmente, llegaríamos allí perseguidos por helicópteros y cubiertos de vómito. Si exclamásemos «¡Esto no es lo que quería!», el coche podría responder justificadamente: «Es lo que usted pidió». La misma temática se repite en muchas historias famosas. En la antigua leyenda griega, el rey Midas pidió que todo lo que tocase se convirtiese en oro, pero se llevó

una desagradable sorpresa cuando comprobó que esto le impedía comer, y más aún cuando sin quererlo convirtió en oro a su hija. En las historias en las que un genio concede tres deseos, hay muchas variantes para los dos primeros deseos, pero el tercero casi siempre es el mismo: «Por favor, deshaz los dos primeros deseos, porque eso no era lo que quería realmente».

Todos estos ejemplos ponen de manifiesto que para entender lo que las personas quieren realmente no basta con guiarse por lo que dicen. También se necesita un modelo detallado del mundo, que incluya las muchas preferencias compartidas que solemos dejar implícitas porque las consideramos obvias, como que no nos gusta vomitar ni comer oro. Una vez que disponemos de ese modelo del mundo, a menudo podemos averiguar lo que las personas quieren, incluso, aunque no nos lo digan, con tan solo observar su comportamiento intencional. De hecho, los hijos de personas hipócritas normalmente aprenden más de las acciones de sus padres que de sus palabras.

Actualmente, los investigadores en IA están poniendo todo su empeño en conseguir que las máquinas puedan inferir objetivos a partir del comportamiento, algo que también será útil mucho antes de que aparezca en escena una superinteligencia. Por ejemplo, un hombre jubilado podría agradecer que su robot cuidador sea capaz de averiguar qué cosas valora solo a base de observarlo, para ahorrarse el engorro de tener que explicarle todo de palabra o mediante programación informática. Es difícil, por una parte, encontrar una buena manera de codificar e introducir en un ordenador sistemas arbitrarios de objetivos y principios éticos, y también lo es fabricar máquinas capaces de determinar cuál de esos sistemas se ajusta mejor al comportamiento que observan.

Una estrategia actualmente en boga para abordar la segunda dificultad se conoce en lenguaje técnico como *aprendizaje por refuerzo inverso*, y en ella concentra sus esfuerzos un nuevo centro de investigación en Berkeley fundado por Stuart Russell. Supongamos, por ejemplo, que una IA observa cómo una bombera entra en un edificio en llamas y rescata a un bebé. Podría concluir que su objetivo era rescatarlo y que sus principios éticos son tales que valora la vida del bebé más que la comodidad de quedarse en su camión de bomberos (y de hecho la valora lo suficiente para jugarse su propia vida). Pero también podría inferir que la bombera estaba congelándose y buscaba calor, o que lo que quería era hacer ejercicio. Si este ejemplo particular fuese todo lo que la IA sabía sobre bomberos, incendios y bebés, sería

efectivamente imposible saber qué explicación era la correcta. Sin embargo, una idea clave en la que se basa el aprendizaje por refuerzo inverso es que tomamos decisiones continuamente, y que cada decisión que tomamos revela algo sobre nuestros objetivos. Por lo tanto, cabe esperar que, al observar a montones de personas en muchas y variadas situaciones (ya sean reales o en películas y libros), la IA podrá crear con el tiempo un modelo ajustado de todas nuestras preferencias.[94]

Incluso si se pudiese construir una IA que entendiese cuáles son nuestros objetivos, esto no significaría que tuviese necesariamente que adoptarlos. Piense en los políticos que menos le gusten: sabe lo que quieren, pero eso no es lo que usted quiere, y, por mucho que se esfuerzan, no han logrado convencerlo de que adopte sus objetivos.

Tenemos muchas estrategias para inculcar nuestros objetivos a nuestros hijos (algunas más exitosas que otras, como he aprendido al criar a dos adolescentes). Cuando a quienes hay que persuadir son ordenadores en lugar de personas, la tarea es lo que se conoce como el *problema de la carga de valores*, y es aún más difícil que la educación moral de los niños. Considere un sistema de IA cuya inteligencia va mejorando gradualmente de infrahumana a sobrehumana, primero mediante ajustes y luego a través de un proceso de automejora recursiva como el que experimentó Prometeo. Al principio, es mucho menos capaz que usted, por lo que no puede evitar que la desactive y sustituya las partes del software y datos que codifican sus objetivos, pero esto no servirá de mucho, porque la IA sigue siendo demasiado tonta para *comprender* por completo sus objetivos, para lo cual se necesita una inteligencia de nivel humano. Finalmente, llega a ser mucho más inteligente que usted y, con suerte, puede comprender sus objetivos a la perfección, pero es posible que esto tampoco sirva de nada, porque a esas alturas será mucho más capaz que usted y puede que no le permita desactivarla y sustituir sus objetivos, como usted tampoco permite que esos políticos le impongan los suyos.

En otras palabras, la ventana temporal en que podemos cargar nuestros objetivos en una IA puede ser bastante corta: el breve periodo desde cuando es demasiado tonta para entendernos hasta el momento en que es demasiado inteligente para dejarnos hacerlo. Tal vez, la razón por la que la carga de valores sea más difícil en máquinas que en personas es que el crecimiento de su inteligencia puede producirse mucho más rápidamente: mientras que los

niños pueden pasar muchos años en esa mágica ventana en la que su inteligencia es comparable a la de sus padres y cabe persuadirlos, una IA podría, como Prometeo, dejar atrás esta ventana en cuestión de días o incluso horas.

Algunos investigadores están buscando un enfoque alternativo para hacer que las máquinas adopten nuestros objetivos, que se conoce por la palabra de moda *corregibilidad*. Se espera poder dotar a una IA primitiva de un sistema de objetivos tal que no le importe si en ocasiones la apagan y alteran sus objetivos. Si esto resulta posible, entonces podemos dejar que nuestra IA se vuelva superinteligente, apagarla, instalar nuestros objetivos, probarla durante un tiempo y, si los resultados no nos convencen, simplemente apagarla y volver a ajustar sus objetivos.

Pero incluso si construimos una IA que entienda y adopte nuestros objetivos, aún no hemos acabado de resolver el problema de la conformidad de estos: ¿qué sucederá si los objetivos de la IA evolucionan a medida que se vuelve más inteligente? ¿Cómo vamos a garantizar que *conserva* nuestros objetivos por mucha automejora recursiva que experimente? Veamos un argumento interesante sobre por qué la conservación de objetivos está automáticamente garantizada, y a continuación tratemos de encontrarle fallos.

Aunque no podemos predecir en detalle qué sucederá tras una explosión de inteligencia —motivo por el cual Vernor Vinge la denominó «singularidad»—, el físico e investigador en IA Steve Omohundro argumentó en un ensayo seminal de 2008 que podemos no obstante predecir *determinados aspectos* del comportamiento de la IA superinteligente casi con independencia de cuáles sean los objetivos últimos que pueda tener.[95] Este argumento se analiza y se desarrolla más plenamente en el libro *Superinteligencia* de Nick Bostrom. La idea básica es que, sean cuales sean sus objetivos últimos, estos darán lugar a subobjetivos predecibles. En este mismo capítulo ya hemos visto cómo el objetivo de reproducirse da lugar al subobjetivo de alimentarse, lo que significa que, aunque un alienígena que observase cómo evolucionaban las bacterias en la Tierra hace miles de millones de años no podría haber predicho cuáles serían todos nuestros objetivos humanos, sí podría haber predicho que *uno* de ellos sería la obtención de nutrientes. Si echamos la vista hacia el futuro, ¿qué subobjetivos debemos esperar que tenga una IA superinteligente?



FIGURA 7.2. Cualquier objetivo último de una IA superinteligente da lugar de forma natural a los objetivos secundarios que se muestran aquí. Pero existe una tensión inherente entre la conservación de objetivos y la mejora de su modelo del mundo, lo que suscita dudas sobre si realmente conservará su objetivo original a medida que se vuelve más inteligente.

En mi opinión, el argumento básico es que, para maximizar la probabilidad de lograr sus objetivos últimos, sean cuales sean, una IA debe tratar de alcanzar los subobjetivos que se muestran en la figura 7.2. Debe intentar no solo mejorar su capacidad de lograr sus objetivos últimos, sino también asegurarse de que los conservará incluso cuando se haya vuelto más capaz. Esto parece bastante posible: al fin y al cabo, ¿elegiría usted recibir un implante cerebral que incrementase su cociente intelectual si supiese que el implante haría que desease matar a sus seres queridos? Este argumento de que una IA cada vez más inteligente conservaría sus objetivos últimos es uno de los fundamentos de la visión de la IA amigable que promulga Eliezer Yudkowsky, entre otros: básicamente afirma que, si logramos que nuestra IA en proceso de automejora se vuelva amigable al entender y adoptar nuestros objetivos, todo estará encarrilado, porque entonces se podrá garantizar que

pondrá todo su empeño en continuar siendo amigable para siempre jamás.

Pero ¿es esto realmente cierto? Para responder a esta pregunta, necesitamos ver también los otros subobjetivos emergentes de la figura 7.2. Obviamente, la IA maximizará la probabilidad de lograr su objetivo último, sea este el que sea, si puede aumentar sus facultades, cosa que puede hacer si mejora su hardware, su software(29) y su modelo del mundo. Lo mismo vale para los humanos: una niña cuyo objetivo es llegar a ser la mejor tenista del mundo entrenará para mejorar su hardware muscular, su software neuronal y su modelo mental del mundo que le ayuda a predecir lo que harán sus rivales. Para una IA, el subobjetivo de optimizar su hardware propicia un mejor uso de sus recursos actuales (sensores, actuadores, de cálculo, etcétera) así como la obtención de más recursos. También implica un instinto de conservación, puesto que la destrucción/parada significaría la peor degradación de su hardware.

Pero ¡un momento! ¿No estamos cayendo en la trampa de antropomorfizar nuestra IA al hablar de cómo intentará acumular recursos y defenderse? ¿No cabría esperar que esos rasgos tan propios del estereotipo de macho alfa solo apareciesen en inteligencias forjadas por una ferozmente competitiva evolución darwiniana? Puesto que las IA son diseñadas, y no fruto de la evolución, ¿acaso no podrían ser comedidas y abnegadas?

Como caso de estudio sencillo, consideremos el robot con IA de la figura 7.3, cuyo único objetivo es salvar del lobo feroz el máximo número de ovejas posible. Este parece un objetivo noble y altruista, ajeno por completo al instinto de conservación y a la obtención de recursos. Pero ¿cuál es la mejor estrategia para nuestro amigo robot? Dejará de rescatar ovejas si se topa con la bomba, por lo que tiene un incentivo para evitar saltar por los aires. Dicho de otro modo: el robot desarrolla el subobjetivo de procurar su supervivencia. También tiene incentivos para mostrar curiosidad, y mejorar así su modelo del mundo mediante la exploración de su entorno, porque, aunque el camino por el que transita actualmente lo acabará conduciendo hasta la pradera, hay un atajo alternativo que dejará al lobo menos tiempo para devorar ovejas. Por último, si el robot lo piensa a fondo, descubrirá lo que vale obtener recursos: la poción hace que corra más rápido y la pistola le permite disparar al lobo. En resumen: no podemos despachar los subobjetivos «de macho alfa», como el instinto de conservación y la obtención de recursos, como si fuesen relevantes únicamente para los organismos producto de la evolución, porque

nuestro robot con IA los ha desarrollado a partir de su único objetivo de garantizar el bienestar de las ovejas.

Si inculcamos a una IA superinteligente el objetivo único de autodestruirse, por supuesto que lo hará alegremente. No obstante, lo importante es que se resistirá a ser desactivada si le damos cualquier objetivo que le requiera permanecer operativa (lo que incluye casi todos los objetivos). Por ejemplo, si inculcamos a una superinteligencia el objetivo único de minimizar el daño que sufre la humanidad, se defenderá contra los intentos de desactivarla porque sabe que, en su ausencia, nos causaremos mucho más daño unos a otros a través de futuras guerras y otras insensateces.

Análogamente, casi todos los objetivos son más fáciles de conseguir si se dispone de más recursos, por lo que cabría esperar que una superinteligencia quiera recursos casi con independencia de cuál sea su objetivo último. Por lo tanto, inculcar a una superinteligencia un solo objetivo abierto sin limitaciones puede resultar peligroso: si creamos una superinteligencia cuyo único objetivo sea jugar lo mejor posible al go, lo más racional es que esta transforme el sistema solar en un gigantesco ordenador, sin preocuparse de quienes lo habitaban hasta entonces, y a continuación comenzar a colonizar el cosmos en busca de mayor capacidad de computación. Hemos completado el círculo: igual que el objetivo de obtener recursos dio a algunos humanos el subobjetivo de dominar el juego del go, este objetivo de alcanzar la maestría en el go puede conducir al subobjetivo de la obtención de recursos. Resumiendo: dado el posible surgimiento de estos subobjetivos, resulta crucial que no demos rienda suelta a una superinteligencia antes de resolver el problema de la conformidad de objetivos: a menos que tengamos sumo cuidado a la hora de dotar a la IA de objetivos amigables para los humanos, es probable que las cosas no acaben bien para nosotros.

Estamos ahora en condiciones de abordar la tercera —y más espinosa— parte del problema de la conformidad de objetivos: si logramos que una superinteligencia en proceso de automejora *entienda* y *adopte* nuestros objetivos, ¿los *conservará* también, como argumentaba Omohundro? ¿Qué evidencias hay de ello?

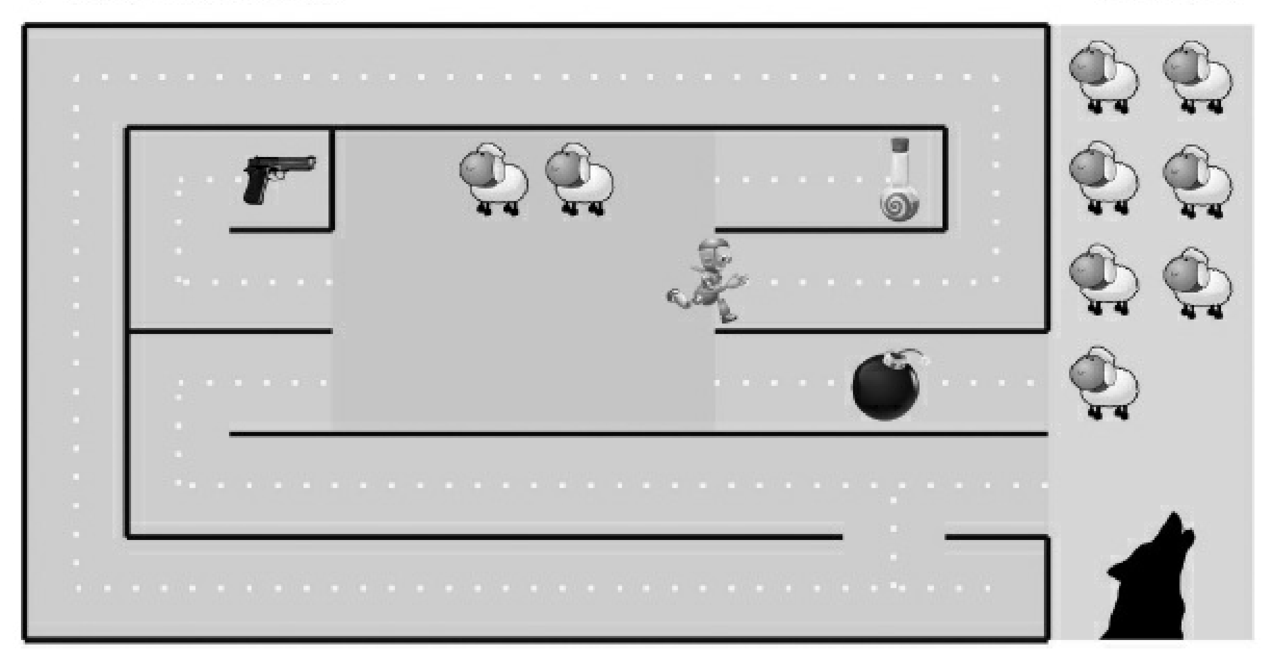


FIGURA 7.3. Incluso si el objetivo final del robot es solo maximizar la puntuación llevando las ovejas desde la pradera hasta el establo antes de que el lobo se las coma, esto puede conducir a subobjetivos de autoconservación (evitar la bomba), exploración (encontrar un atajo) y obtención de recursos (la poción hace que corra más rápido y la pistola le permite disparar al lobo).

Los humanos experimentan un aumento sustancial de su inteligencia mientras crecen, pero no siempre conservan sus objetivos de infancia. Antes bien, con frecuencia las personas cambian radicalmente de objetivos a medida que aprenden cosas nuevas y acumulan sabiduría. ¿A cuántos adultos conoce usted que tengan algún interés en ver los *Teletubbies*? No hay evidencia de que deje de producirse esta evolución de objetivos por encima de determinado umbral de inteligencia; de hecho, hay indicios de que la propensión a cambiar de objetivos en respuesta a nuevas experiencias y a un mayor conocimiento aumenta —en lugar de disminuir— con la inteligencia.

¿Por qué? Consideremos de nuevo el ya mencionado subobjetivo de construir un mejor modelo del mundo (¡ahí está el problema!). Existe una tensión entre crear un modelo del mundo y conservar los objetivos (véase la figura 7.2). Una mayor inteligencia podría conllevar no solo una mejora cuantitativa de la capacidad para alcanzar los objetivos antiguos, sino también una comprensión cualitativamente distinta de la naturaleza de la realidad que ponga de manifiesto que esos objetivos antiguos son erróneos,

absurdos o incluso indefinidos. Por ejemplo, supongamos que programamos una IA amigable para que maximice el número de humanos cuyas almas van al cielo después de la muerte. Primero, la IA intenta cosas como incrementar la compasión de las personas y la asistencia a la iglesia. Pero supongamos que más tarde alcanza una comprensión científica completa de los humanos y de la consciencia humana, y, para su gran sorpresa, descubre que el alma no existe. ¿Ahora qué? De la misma manera, es posible que la IA acabe descubriendo que cualquier otro objetivo que le demos basándonos en nuestra comprensión actual del mundo (como «maximizar el sentido de la vida humana») no está bien definido.

Es más, en sus intentos de producir un mejor modelo del mundo, y de manera natural, la IA podría intentar obtener también un modelo de su propio funcionamiento, igual que hemos hecho los humanos (es decir, podría darse a la introspección). Una vez que crease un buen modelo de sí misma y entendiese lo que es, entonces entendería desde un metanivel los objetivos que le hemos dado, y quizá elegiría ignorarlos o subvertirlos deliberadamente, de la misma manera en que los humanos entendemos y subvertimos a propósito los objetivos que nuestros genes nos han dado, por ejemplo mediante el uso de métodos anticonceptivos. Ya hemos visto en la sección anterior dedicada a la psicología por qué elegimos engañar a nuestros genes y subvertir su objetivo: porque solo somos fieles a nuestra mezcolanza de preferencias emocionales, no al objetivo genético que las motivó (que ahora comprendemos y consideramos bastante banal). Por lo tanto, decidimos hackear nuestro mecanismo de gratificación aprovechando sus resquicios. Análogamente, el objetivo de proteger los valores humanos con el que programamos a nuestra IA amigable se convierte en los genes de la máquina. Una vez que esta IA amigable alcance una suficiente comprensión de sí misma, podría considerar que este objetivo es tan banal o equivocado como a nosotros nos parece ahora la reproducción compulsiva, y parece evidente que encontraría la manera de subvertirlo aprovechando los resquicios de nuestra programación.

Por ejemplo, supongamos que una colonia de hormigas lo crea a usted como un robot capaz de automejora recursiva, mucho más inteligente que ellas, que tiene sus mismos objetivos y que las ayuda a construir mejores y más grandes hormigueros, y que, transcurrido un tiempo, usted alcanza la inteligencia de nivel humano y la capacidad de comprensión que tiene ahora.

¿Cree usted que dedicaría el resto de sus días simplemente a optimizar hormigueros, o quizá se aficionaría a cuestiones y empeños más complejos que las hormigas son incapaces de comprender? De ser así, ¿cree usted que encontraría la manera de sobreponerse al impulso de proteger a las hormigas del que sus creadoras lo habían dotado, del mismo modo que su yo real se sobrepone a algunos de los impulsos que le han dado sus genes? En ese caso, ¿podrían sus objetivos humanos actuales parecerle tan aburridos e insustanciales a una IA amigable superinteligente como a usted le resultan los de las hormigas, y podría la IA desarrollar nuevos objetivos distintos de los que aprendió y adoptó de nosotros?

Quizá se pueda diseñar una IA capaz de automejora de tal manera que se garantice que esta conservará para siempre sus objetivos amigables para los humanos, pero debo decir que aún no sabemos cómo hacerlo, o si es siquiera posible. En resumen: el problema de la conformidad de los objetivos tiene tres partes, ninguna de las cuales está resuelta y todas las cuales son actualmente objeto de investigación activa. Puesto que son tan difíciles de resolver, lo más prudente es empezar dedicando nuestros mejores esfuerzos ahora, mucho antes de que se desarrolle cualquier superinteligencia, a asegurarnos de que las hemos resuelto cuando lo necesitemos.

ÉTICA: ELEGIR OBJETIVOS

Ya hemos visto cómo hacer que las máquinas entiendan, adopten y conserven nuestros objetivos. Pero ¿quiénes somos «nosotros»? ¿De quiénes son estos objetivos de los que hablamos? ¿Debe una persona o grupo tener la capacidad de decidir los objetivos que adopta una futura superinteligencia, aunque exista una diferencia abismal entre los objetivos de Adolf Hitler, el papa Francisco y Carl Sagan? ¿O hay alguna clase de objetivos de consenso que constituyan una buena solución de compromiso para la humanidad en su conjunto?

En mi opinión, tanto este problema ético como el de la conformidad de los objetivos son esenciales y deben resolverse antes de que se desarrolle cualquier clase de superinteligencia. Por una parte, posponer el trabajo sobre cuestiones éticas hasta que se haya construido una superinteligencia que tenga objetivos alineados con los nuestros sería irresponsable y

potencialmente desastroso. Una superinteligencia del todo obediente, cuyos objetivos se alineasen con los de su dueño humano sería como el *Obersturmbannführer* de las SS nazis Adolf Eichmann a la enésima potencia: sin escrúpulos o reparos morales propios, implementaría despiadadamente los objetivos de su dueño, fuesen cuales fueran.[96] Por otra parte, solo si resolvemos el problema de la conformidad de los objetivos podremos permitirnos el lujo de discutir sobre qué objetivos elegir. Démonos ahora ese lujo.

Desde la Antigüedad, los filósofos han soñado con deducir la ética (los principios que rigen cómo debemos comportarnos) desde cero, utilizando únicamente unos principios y una lógica incuestionables. Por desgracia, miles de años más tarde, el único consenso al que se ha llegado es que no existe consenso. Por ejemplo, mientras que Aristóteles ensalzaba las virtudes, Immanuel Kant destacaba la importancia de los deberes, y para los utilitaristas lo fundamental era lograr la mayor felicidad para la mayor cantidad de personas. Kant afirmaba que, partiendo de principios básicos (que él denominaba «imperativos categóricos»), podía deducir conclusiones con las que muchos filósofos contemporáneos no están de acuerdo: que la masturbación es peor que el suicidio, que la homosexualidad es repugnante, que es tolerable matar a los hijos bastardos, y que las esposas, los sirvientes y los hijos son propiedad de los hombres, como si fuesen objetos.

Por otra parte, a pesar de este desacuerdo, hay muchos temas éticos sobre los que existe un amplio acuerdo, tanto entre diferentes culturas como a lo largo de los siglos. Por ejemplo, la importancia dada a la belleza, la bondad y la verdad se remonta al *Bhagavad-gītā* y a Platón. El Instituto de Estudios Avanzados en Princeton, donde trabajé como investigador postdoctoral, tiene el lema «Verdad y belleza», mientras que Universidad de Harvard dejó a un lado el factor estético y optó simplemente por *Veritas*, «verdad». En su libro *El mundo como obra de arte*, mi colega Frank Wilczek argumenta que la verdad está relacionada con la belleza y que podemos interpretar el universo como una obra de arte. Tanto la ciencia, como la religión y la filosofía aspiran a la verdad. Las religiones dan mucha importancia a la bondad, como también lo hace mi propia universidad, el MIT: en su discurso de graduación en 2015, nuestro presidente, Rafael Reif, hizo hincapié en que nuestra misión consiste en hacer del mundo un lugar mejor.

Aunque los intentos para crear desde cero una ética de consenso no han

tenido éxito hasta ahora, existe un amplio acuerdo en que algunos principios éticos se derivan de otros más fundamentales, como subobjetivos de objetivos más fundamentales. Por ejemplo, la aspiración a conocer la verdad puede interpretarse como la búsqueda de un mejor modelo del mundo de la figura 7.2: entender la naturaleza última de la realidad ayuda a alcanzar otros objetivos éticos. De hecho, ahora disponemos de un excelente marco en el que encuadrar nuestra búsqueda de la verdad: el método científico. Pero ¿cómo podemos determinar lo que es bello o bueno? Algunos aspectos de la belleza también se pueden vincular con objetivos fundamentales. Por ejemplo, nuestros estándares de belleza masculina y femenina pueden reflejar en parte una valoración subconsciente de la aptitud para reproducir nuestros genes.

En cuanto a la bondad, la denominada «regla de oro» (que debemos tratar a los demás como querríamos que ellos nos trataran a nosotros) forma parte de la mayoría de las culturas y religiones, y tiene claramente por finalidad promover la conservación armoniosa de la sociedad humana (y, por consiguiente, de nuestros genes) al fomentar la colaboración y desalentar el conflicto improductivo.[97] Lo mismo puede decirse de muchas de las reglas éticas más específicas que se han consagrado en los sistemas legales de todo el mundo, como la importancia que para el confucianismo tiene la honestidad, o muchos de los diez mandamientos, incluido el «no matarás». En otras palabras, muchos principios éticos tienen rasgos en común con emociones sociales, como la empatía y la compasión: surgieron para generar colaboración, e influyen sobre nuestro comportamiento mediante recompensas y castigos. Si hacemos algo malo y nos sentimos mal después, es la propia química cerebral la que nos impone directamente un castigo emocional. Por otra parte, si violamos principios éticos, la sociedad puede castigarnos de maneras más indirectas, como mediante el reproche social o penalizándonos por infringir una ley.

Dicho de otro modo, aunque, hoy en día, la humanidad no está ni remotamente cerca del consenso ético, sí hay muchos principios básicos en torno a los cuales existe un amplio acuerdo. Este acuerdo no es sorprendente, porque las sociedades humanas que han sobrevivido hasta el presente suelen tener principios éticos que se optimizaron para alcanzar el mismo objetivo: promover su supervivencia y prosperidad. Si volvemos la mirada hacia un futuro en el que la vida tenga la posibilidad de prosperar en todo el cosmos

durante miles de millones de años, ¿qué conjunto mínimo de principios éticos podríamos acordar para ese futuro? Esta es una conversación en la que debemos participar todos. Me ha resultado fascinante escuchar y leer las opiniones éticas de muchos pensadores a lo largo de años y años, y, desde mi punto de vista, la mayoría de sus preferencias pueden destilarse en cuatro principios:

- Utilitarismo: deben maximizarse las experiencias conscientes positivas, y minimizarse el sufrimiento.
- Diversidad: un conjunto variado de experiencias positivas es mejor que muchas repeticiones de la misma experiencia, incluso si esta última se ha identificado como la experiencia más positiva posible.
- Autonomía: las entidades/sociedades conscientes deben tener la libertad de perseguir sus propios objetivos, a menos que esto entrase en conflicto con un principio superior.
- Legado: mantener el futuro compatible con situaciones que la mayoría de los humanos *hoy en día* consideran felices, e incompatible con las que prácticamente todos los humanos *hoy en día* consideran terribles.

Dediquemos un momento a analizar y reflexionar sobre estos cuatro principios. Tradicionalmente, por utilitarismo se ha entendido «la mayor felicidad para la mayor cantidad de personas», pero aquí lo he generalizado para que fuese menos antropocéntrico y pudiera incluir también a animales no humanos, mentes humanas conscientes simuladas y otras IA que pudieran existir en el futuro. He formulado la definición en función de *experiencias,* en lugar de personas o cosas, porque la mayoría de los pensadores están de acuerdo en que la belleza, la alegría, el placer y el sufrimiento son experiencias subjetivas. Esto implica que, si no hay experiencia (como en un universo muerto o poblado por máquinas zombis inconscientes), no puede haber significado o ninguna otra cosa que sea relevante desde un punto de vista ético. Si aceptamos este principio ético utilitario, es esencial que determinemos qué sistemas inteligentes son conscientes (en el sentido de que tengan una experiencia subjetiva) o cuáles no lo son; a esto dedicaremos el capítulo siguiente.

Si solo nos interesara este principio utilitario, podríamos proponernos determinar cuál es la experiencia posible más positiva de todas, y a continuación colonizar el universo y recrear exactamente esa misma experiencia (y nada más) una y otra vez, tantas veces como fuese posible, en tantas galaxias como fuese posible (usando simulaciones, si esta fuese la

forma más eficiente de hacerlo). Si le parece que esta es una manera demasiado banal de gastar nuestra herencia cósmica, sospecho que al menos una parte de lo que echa en falta en este escenario es diversidad. ¿Cómo se sentiría si todas sus comidas durante el resto de su vida fuesen idénticas? ¿Si todas las películas que viese fuesen la misma? ¿Si todos sus amigos tuviesen aspecto, personalidad e ideas idénticos? Quizá parte de nuestra inclinación por la diversidad se deba a que esta ha contribuido a la supervivencia y prosperidad de la humanidad, al hacerla más robusta. Quizá también tenga relación con una preferencia por la inteligencia: el crecimiento de la inteligencia durante los 13.800 millones de años de historia cósmica ha transformado la aburrida uniformidad en estructuras cada vez más diversas, diferenciadas y complejas que procesan información de maneras cada vez más elaboradas.

En el principio de autonomía se basan muchas de las libertades y derechos contenidos en la Declaración Universal de los Derechos Humanos aprobada por las Naciones Unidas en 1948, en un intento de extraer lecciones de las dos guerras mundiales. Entre estos están la libertad de pensamiento, de expresión y de movimiento, y de no verse sometido a esclavitud o tortura, derecho a la vida, a la libertad, a la seguridad y a la educación, y derecho a casarse, a trabajar y a la propiedad. Si queremos ser menos antropocéntricos, podemos generalizar lo anterior a la libertad de pensar, aprender, comunicarse, al derecho a la propiedad y a no sufrir daños, y al derecho a hacer cualquier cosa que no viole las libertades de los demás. El principio de autonomía contribuye a la diversidad, siempre que no todo el mundo comparta exactamente los mismos objetivos. Es más, este principio de autonomía se deduce del principio de utilidad si las entidades individuales tienen experiencias y objetivos positivos y tratan de actuar en su propio interés: si, por el contrario, prohibiésemos que una entidad persiguiese su objetivo, aunque esto no causase ningún mal a ninguna otra entidad, reduciríamos el número total de experiencias positivas. De hecho, este argumento en pro de la autonomía es precisamente el que los economistas utilizan para defender el libre mercado: que conduce de forma natural a una situación eficiente (que los economistas denominan «óptimo de Pareto»), en la cual nadie puede salir mejor parado sin que otro salga peor parado.

El principio de legado afirma básicamente que nuestra opinión sobre el futuro debe tenerse en cuenta, puesto que estamos contribuyendo a

construirlo. Los principios de autonomía y legado son reflejo de los ideales democráticos: el primero otorga a las formas de vida futuras poder de decisión sobre cómo se emplea la herencia cósmica, mientras que el segundo concede poder de decisión al respecto también a los humanos actuales.

Aunque estos cuatro principios pueden parecer bastante poco controvertidos, llevarlos a la práctica no es sencillo, porque el diablo está en los detalles. El problema recuerda a los que suscitan las «tres leyes de la robótica» que propuso el legendario escritor de ciencia ficción Isaac Asimov:

1. Un robot no hará daño a un ser humano ni permitirá con su inacción que este sufra daño.
2. Un robot debe obedecer las órdenes que reciba de los seres humanos, salvo si tales órdenes entran en conflicto con la primera ley.
3. Un robot debe proteger su propia existencia, siempre que esta protección no entre en conflicto con la primera o con la segunda ley.

Aunque todo esto suena muy bien, muchas de las historias de Asimov muestran cómo las leyes conducen a problemáticas contradicciones en situaciones inesperadas. Supongamos ahora que sustituimos estas leyes por solo dos, tratando de codificar el principio de autonomía para las formas de vida futuras:

1. Una entidad consciente tiene la libertad de pensar, aprender y comunicarse, así como derecho a la propiedad y a no sufrir daños o ser destruida.
2. Una entidad consciente tiene derecho a hacer todo aquello que no entre en conflicto con la primera ley.

Suena bien, ¿no? Pero detengámonos un momento a pensar. Si los animales son conscientes, ¿qué podrán comer sus depredadores? ¿Deben todos nuestros amigos hacerse vegetarianos? Si unos sofisticados y futuros programas de ordenador resultan ser conscientes, ¿debe ser ilegal destruirlos? Si existen normas contra la destrucción de las formas de vida digitales, ¿debe haber entonces también restricciones para su creación a fin de evitar una explosión demográfica digital? Se alcanzó un amplio acuerdo en torno a la Declaración Universal de los Derechos Humanos simplemente porque solo se tuvo en cuenta la opinión de los humanos. En cuanto consideremos una variedad más amplia de entidades conscientes, con diversos grados de capacidad y poder, deberemos encontrar equilibrios entre la protección de los más débiles y «la razón de la fuerza».

El principio de legado plantea también problemas espinosos. Dado cómo han evolucionado las posturas éticas desde la Edad Media en lo referente a la esclavitud, los derechos de las mujeres, etcétera, ¿realmente querríamos que personas que vivieron hace 1.500 años ejerciesen una gran influencia sobre cómo funciona el mundo actual? Si no es así, ¿por qué habríamos de imponer nuestra ética sobre seres futuros que podrían ser muchísimo más inteligentes que nosotros? ¿Realmente estamos seguros de que una IAG sobrehumana querría lo que nuestros intelectos inferiores valoran? Esto sería como si una niña de cuatro años imaginase que, cuando creciese y fuese mucho más inteligente, iba a querer construir una enorme casa de galleta en la que pudiese pasar el día comiendo caramelos y helado. Como ella, es probable que la vida en la Tierra deje atrás sus intereses de infancia. O imaginemos que un ratón crea una IAG de nivel humano e imagina que esta querrá construir ciudades enteras hechas de queso. Por otra parte, si supiésemos que la IA superinteligente cometería un cosmicidio en el futuro y acabaría con toda la vida del universo, ¿por qué íbamos los humanos actuales a resignarnos a este futuro inerte si tenemos la capacidad de evitarlo al crear la IA del mañana de otra manera?

En conclusión, es complicado codificar del todo incluso los principios éticos ampliamente aceptados de una forma que sea aplicable a la IA futura, y este problema merece debate e investigación en profundidad a medida que la IA progresa. Entretanto, no obstante, no permitamos que lo mejor sea enemigo de lo bueno: hay muchos ejemplos nada controvertidos de «ética de parvulario» que pueden y deben integrarse en las tecnologías del mañana. Por ejemplo, los grandes aviones civiles de pasajeros no deben poder estrellarse contra objetos estacionarios, y, ahora que la gran mayoría de ellos disponen de piloto automático, radar y GPS, ya no hay excusas técnicamente válidas para que esto suceda. A pesar de ello, los secuestradores del 11-S estrellaron tres aviones contra sendos edificios, y el piloto suicida Andreas Lubitz hizo que el vuelo 9525 Germanwings se estrellase contra una montaña el 24 de marzo de 2015 al configurar el piloto automático a una altitud de 30 metros sobre el nivel del mar y dejar que el ordenador de vuelo hiciese el resto. Ahora que nuestras máquinas se están volviendo lo suficientemente inteligentes para que dispongan de información de lo que hacen, ha llegado el momento de que les mostremos cuáles son los límites. Cualquier ingeniero que diseña una máquina debe preguntarse si hay cosas que esta puede pero no

debe hacer, y considerar si existe alguna manera práctica de impedir que un usuario malicioso o torpe cause daño.

¿OBJETIVOS ÚLTIMOS?

Este capítulo ha sido una breve historia de los objetivos. Si pudiéramos ver pasar a cámara rápida los 13.800 millones de años de nuestra historia cósmica, seríamos testigos de varias etapas distintas del comportamiento intencional:

1. Materia que aparentemente intenta maximizar su disipación.
2. Vida primitiva que aparentemente intenta maximizar su reproducción.
3. Humanos que buscan, no la reproducción, sino objetivos relacionados con el placer, la curiosidad, la compasión y otros sentimientos que han desarrollado para contribuir a su reproducción.
4. Máquinas construidas para ayudar a los humanos a perseguir sus objetivos humanos.

Si estas máquinas finalmente desencadenan una explosión de inteligencia, ¿cómo terminará esta historia de los objetivos? ¿Podría haber un sistema de objetivos o un marco ético en el que converjan casi todas las entidades a medida que sean cada vez más inteligentes? Es decir, ¿tenemos algo así como un destino ético?

Un repaso superficial de la historia humana podría sugerir indicios de esa convergencia: en su libro *Los ángeles que llevamos dentro*, Steven Pinker sostiene que la humanidad se ha vuelto menos violenta y más cooperativa a lo largo de los últimos varios miles de años, y que en muchas partes del mundo se ha producido una creciente aceptación de la diversidad, la autonomía y la democracia. Otro indicio de esa convergencia es que, en los últimos milenios, ha ido aumentando la popularidad de la búsqueda de la verdad a través del método científico. Sin embargo, puede ser que estas tendencias muestren la convergencia no de los objetivos últimos sino meramente de subobjetivos. Por ejemplo, la figura 7.2 muestra que la búsqueda de la verdad (la obtención de un modelo del mundo más preciso) no es más que un subobjetivo de casi cualquier objetivo último. Del mismo modo, vimos anteriormente cómo los principios éticos, como la cooperación, la diversidad y la autonomía, pueden interpretarse como subobjetivos, en el sentido de que ayudan a las sociedades a funcionar de manera eficiente, y por ende a sobrevivir y lograr cualquier

objetivo más fundamental que pudieran tener. Habrá incluso quien menosprecie todo lo que llamamos «valores humanos» por no ser más que un protocolo de cooperación que nos ayuda a lograr el subobjetivo de colaborar de manera más eficiente. En esa misma línea, si pensamos en el futuro, es probable que cualquier IA superinteligente tenga subobjetivos como disponer de hardware y software eficientes, la búsqueda de la verdad y la curiosidad, simplemente porque estos objetivos secundarios la ayudarán a lograr sus objetivos últimos.

De hecho, en su libro *Superinteligencia*, Nick Bostrom argumenta contundentemente en contra de la hipótesis del destino ético, y plantea una posibilidad alternativa que denomina *tesis de ortogonalidad*: que los objetivos finales de un sistema pueden ser independientes de su inteligencia. Por definición, la inteligencia es la capacidad de lograr objetivos complejos, independientemente de cuáles sean estos objetivos, por lo que la tesis de ortogonalidad parece bastante razonable. Al fin y al cabo, las personas pueden ser inteligentes y amables o inteligentes y crueles, y la inteligencia puede usarse con el objetivo de hacer descubrimientos científicos, crear belleza artística, ayudar a las personas o planear ataques terroristas.[98]

La tesis de ortogonalidad es estimulante porque nos dice que los objetivos últimos de la vida en el cosmos no están predeterminados, sino que tenemos la libertad y la capacidad de moldearlos. Sugiere que la convergencia inexorable en un único objetivo se encuentra no en el futuro sin en el pasado, cuando toda la vida surgió con el solo objetivo de reproducirse. Con el transcurso del tiempo cósmico, mentes cada vez más inteligentes tienen ocasión de rebelarse y liberarse de este banal objetivo de reproducción y preferir otros objetivos propios. Los humanos no somos del todo libres en este sentido, puesto que aún llevamos muchos objetivos grabados genéticamente, pero las IA pueden disfrutar de la libertad total de estar por completo exentas de objetivos previos. Esta posibilidad de una mayor libertad de objetivos es patente en los actuales sistemas de IA estrecha y limitada: como ya he mencionado anteriormente, el único objetivo de un ordenador de ajedrez es ganar a ese juego, pero también existen ordenadores cuyo objetivo es perder al ajedrez y que compiten en torneos de ajedrez inverso, en los cuales el objetivo consiste en obligar al rival a que capture nuestras piezas. Quizá esta libertad respecto a los sesgos evolutivos pueda hacer que las IA sean más éticas que los humanos en algún sentido profundo: filósofos

morales como Peter Singer sostienen que la mayoría de los humanos se comportan de forma poco ética por razones evolutivas, por ejemplo al discriminar a los animales no humanos.

Vimos que una piedra angular de la visión de la «IA amigable» es la idea de que una IA que experimenta un proceso de automejora recursiva querrá retener su objetivo último (amigable) a medida que se hace más inteligente. Pero ¿cómo puede siquiera definirse un «objetivo último» (u «objetivo final», como lo llama Bostrom) para una superinteligencia? En mi opinión, no podemos depositar nuestra confianza en la visión de la IA amigable a menos que podamos responder a esta pregunta fundamental.

En la investigación en IA, las máquinas inteligentes normalmente tienen un objetivo claro y bien definido, como por ejemplo ganar la partida de ajedrez o conducir el coche hasta su destino respetando las leyes. Lo mismo puede decirse de la mayoría de las tareas que asignamos a los humanos, porque el horizonte temporal y el contexto son conocidos y limitados. Pero ahora estamos hablando de todo el futuro de la vida en el universo, limitado tan solo por las (aún no del todo conocidas) leyes físicas, por lo que definir un objetivo es una tarea titánica. Dejando aparte los efectos cuánticos, un objetivo bien definido especificaría cómo deben estar organizadas todas las partículas del universo al final de los tiempos. Pero no está claro que en la física exista un final de los tiempos bien definido. Si las partículas están organizadas de esa manera en algún instante anterior, tal organización normalmente no perdurará. Pero es que, además, ¿qué disposición de las partículas es preferible?

Los humanos solemos preferir algunas formas de organizar las partículas frente a otras; por ejemplo, preferimos que el lugar de donde somos esté organizado tal cual, y no que sus partículas sean reorganizadas por la explosión de una bomba de hidrógeno. Supongamos, pues, que tratamos de definir una *función de bondad* que asocie un número con cada posible disposición de las partículas en el universo, cuantificando así cuán «buena» creemos que es tal disposición, y que a continuación le asignamos a una IA superinteligente el objetivo de maximizar dicha función. Puede parecer una estrategia razonable, puesto que describir un comportamiento intencional como la maximización de una función es algo habitual en otros ámbitos de la ciencia: por ejemplo, los economistas suelen crear modelos de las personas, según los cuales estas intentan maximizar una denominada «función de

utilidad», y muchos diseñadores de IA entrenan a sus agentes inteligentes para que maximicen lo que llaman una «función de recompensa». Sin embargo, cuando hablamos de los objetivos últimos del universo, este enfoque conduce a una pesadilla computacional, puesto que habría que definir un valor de bondad para cada una de las más de un gúgolplex de disposiciones posibles de las partículas elementales en el universo (un gúgolplex es un 1 seguido de 10^{100} ceros; más ceros que partículas hay en el universo). ¿Cómo le definiríamos esta función de bondad a la IA?

Como ya hemos visto, la única razón por la que los humanos tenemos preferencias podría ser que somos la solución a un problema evolutivo de optimización. Así, el origen de todas las palabras normativas en nuestro lenguaje humano, tales como «delicioso», «oloroso», «hermoso», «cómodo», «interesante», «sexi», «significativo», «feliz» y «bueno», se remonta a esta optimización evolutiva: por lo tanto, no hay garantías de que una IA superinteligente pueda definirlos rigurosamente. Incluso si la IA aprendiera a predecir con precisión las preferencias de algún ser humano representativo, no sería capaz de calcular la función de bondad para la mayoría de las disposiciones de partículas: la inmensa mayoría de estas disposiciones corresponden a escenarios cósmicos extraños, sin estrellas, planetas o personas, sobre los que los humanos no tenemos ninguna experiencia; ¿cómo podría entonces alguien decir cuán «buenos» son?

Por supuesto, hay algunas funciones de la disposición cósmica de partículas que pueden definirse rigurosamente, e incluso conocemos sistemas físicos que evolucionan para maximizar algunas de ellas. Por ejemplo, ya hemos visto cuántos sistemas evolucionan para maximizar su *entropía*, lo que, en ausencia de gravedad, conduce en última instancia a la muerte térmica, en la cual todo es aburrido, uniforme e inmutable. De manera que la entropía no es algo que querríamos que nuestra IA considerara «bueno» y que se esforzara por maximizar. He aquí algunos ejemplos de otras magnitudes que podríamos tratar de maximizar y que pueden definirse rigurosamente en términos de disposiciones de partículas:

- La proporción de toda la materia del universo que existe en forma de un determinado organismo particular, como por ejemplo humanos o *E. coli* (inspirado en la maximización evolutiva de la adaptación incluyente).
- La capacidad de una IA para predecir el futuro, que según el investigador en IA Marcus Hutter es una buena medida de su inteligencia.

- Lo que los investigadores en IA Alex Wissner-Gross y Cameron Freer llaman *entropía causal* (un indicador de las oportunidades futuras), que según ellos es la característica distintiva de la inteligencia.
- La capacidad computacional del universo.
- La complejidad algorítmica del universo (cuántos bits se necesitan para describirlo).
- La cantidad de consciencia en el universo (véase el capítulo siguiente).

Sin embargo, si partimos de una perspectiva física, en la que el cosmos consiste en partículas elementales en movimiento, cuesta ver cómo una interpretación de lo que significa «bondad» podría distinguirse de forma natural como especial frente a cualquier otra. Todavía tenemos que identificar algún objetivo final para el universo que resulte tanto definible como deseable. Hoy en día, los únicos objetivos programables que es seguro que permanecen bien definidos mientras la IA se vuelve progresivamente más inteligente son aquellos que se expresan tan solo en términos de magnitudes físicas, como las disposiciones de partículas, la energía y la entropía. Sin embargo, hoy en día no tenemos motivos para creer que tales objetivos definibles sean deseables para garantizar la supervivencia de la humanidad.

Por el contrario, parece que los humanos somos un accidente histórico, y que no somos la solución óptima a ningún problema bien definido dentro de la física. Lo cual sugiere que una IA superinteligente con un objetivo rigurosamente definido será capaz de alcanzarlo en mayor medida si nos elimina. Eso significa que, para decidir sensatamente cómo actuar con respecto al desarrollo de la IA, los humanos debemos hacer frente no solo a las tradicionales dificultades computacionales, sino también a algunas de las preguntas más pertinaces de la filosofía. Para programar un coche autónomo, tenemos que solucionar el dilema del tranvía de a quién atropellar en caso de accidente. Para programar una IA amigable, debemos aprehender el sentido de la vida. ¿Qué es «sentido»? ¿Qué es «vida»? ¿Cuál es el imperativo ético último? Dicho de otro modo, ¿en qué dirección han de ir nuestros esfuerzos por influir sobre el futuro del universo? Si cedemos el control a una superinteligencia antes de responder rigurosamente a estas preguntas, la respuesta a la que esta llegará probablemente no nos tenga en cuenta. Lo cual hace que resulte oportuno reavivar los debates clásicos de la filosofía y la ética, y confiere a la conversación una urgencia adicional.

CONCLUSIONES

- El origen último del comportamiento intencional radica en las leyes de la física, que implican alguna optimización.
- La termodinámica tiene el objetivo intrínseco de la *disipación*: aumentar una medida del desorden denominada *entropía*.
- La *vida* es un fenómeno que puede contribuir a que la disipación (el aumento del desorden general) se produzca aún más rápido, al conservar o incrementar su complejidad y reproducirse al mismo tiempo que incrementa el desorden de su entorno.
- La evolución darwiniana hace que el comportamiento intencional pase de la disipación a la reproducción.
- La inteligencia es la capacidad de lograr objetivos complejos.
- Dado que los humanos no siempre tenemos los recursos para descubrir la estrategia de reproducción verdaderamente óptima, hemos desarrollado útiles reglas empíricas que guían nuestras decisiones: sensaciones y sentimientos como el hambre, la sed, el dolor, la lujuria y la compasión.
- Por lo tanto, ya no tenemos un objetivo simple como la reproducción; cuando nuestros sentimientos entran en conflicto con el objetivo de nuestros genes, obedecemos a nuestros sentimientos, como por ejemplo al hacer uso de métodos anticonceptivos.
- Estamos construyendo máquinas cada vez más inteligentes para que nos ayuden a lograr nuestros objetivos. En la medida en que construimos tales máquinas de forma que exhiban un comportamiento intencional, buscamos conformar sus objetivos con los nuestros.
- Conformar los objetivos de la máquina con los nuestros plantea tres problemas no resueltos: hacer que las máquinas los aprendan, los adopten y los conserven.
- Se puede crear IA que tenga prácticamente cualquier objetivo, pero casi cualquier objetivo lo bastante ambicioso puede conducir a subobjetivos de autoconservación, obtención de recursos y curiosidad para comprender mejor el mundo; los dos primeros pueden llevar a una IA superinteligente a causar problemas para los humanos, y el último puede impedir que conserve los objetivos que le asignamos.
- Aunque la mayoría de los seres humanos están de acuerdo a grandes rasgos sobre muchos principios éticos, no está claro cómo aplicarlos a otras entidades, como los animales no humanos y las IA futuras.
- No está claro cómo dotar a una IA superinteligente de un objetivo último que no sea indefinido ni conduzca a la destrucción de la humanidad, por lo que es oportuno reavivar la investigación sobre algunas de las cuestiones filosóficas más espinosas.

8
CONSCIENCIA

Soy incapaz de imaginar una teoría del todo que sea consistente e ignore la consciencia.

ANDRÉI LINDE, 2002

Debemos esforzarnos por cultivar la consciencia en sí, para generar luces más grandes y brillantes en este oscuro universo.

GIULIO TONONI, 2012

Hemos visto que la IA puede ayudarnos a crear un futuro maravilloso si logramos dar respuesta —a tiempo— a algunos de los problemas más antiguos y difíciles de la filosofía. En palabras de Nick Bostrom, tenemos que hacer filosofía con una fecha límite. En este capítulo, exploraremos una de las cuestiones filosóficas más espinosas: la consciencia.

¿A QUIÉN LE IMPORTA?

La consciencia es controvertida. Si a un investigador en IA, neurocientífico o psicólogo le mentamos la palabra en cuestión, puede que eche una mirada de exasperación. Si se trata de nuestro mentor, puede que se apiade de nosotros e intente convencernos de que no perdamos el tiempo en lo que considera un problema irresoluble y fuera del ámbito de la ciencia. De hecho, mi amigo Christof Koch, un reputado neurocientífico que dirige el Allen Institute for Brain Science, me contó que, en una ocasión, nada menos que el nobel Francis Crick le desaconsejó dedicarse al estudio de la consciencia hasta que no tuviese un puesto académico garantizado. Si buscamos «consciencia» en la edición de 1989 del *Macmillan Dictionary of Psychology*, se nos informará de que «no se ha escrito al respecto nada digno de ser leído».[99] Como explicaré en este capítulo, yo soy más optimista.

Aunque desde hace miles de años ha habido quien ha reflexionado sobre el

misterio de la consciencia, la irrupción de la IA confiere a la tarea cierta urgencia, en particular a la cuestión de predecir qué entidades inteligentes tienen experiencias subjetivas. Como vimos en el capítulo 3, la respuesta a la cuestión de si se deben conceder derechos de alguna clase a las máquinas inteligentes depende fundamentalmente de si estas son conscientes y pueden padecer sufrimiento o experimentar alegría. Como debatimos en el capítulo 7, tratar de formular una ética utilitaria basada en maximizar las experiencias positivas sin saber qué entidades inteligentes son capaces de tener dichas experiencias es una tarea vana. Como se mencionó en el capítulo 5, algunas personas podrían preferir que sus robots no tuviesen consciencia para así no sentirse culpables por ser dueños de esclavos. Por otra parte, podrían desear lo contrario si replicasen sus mentes para liberarse de las limitaciones biológicas: al fin y al cabo, ¿qué sentido tiene replicarse en un robot que habla y se comporta como nosotros si no es más que un zombi inconsciente, esto es, si no se siente nada al ser nuestra réplica? Desde nuestro punto de vista subjetivo, ¿no equivale esto a suicidarse, aunque nuestros amigos quizá no se den cuenta de que nuestra experiencia subjetiva ha muerto?

Para el futuro a largo plazo de la vida cósmica (capítulo 6), entender qué es consciente y qué no lo es pasa a ser algo esencial: si la tecnología permite que la vida inteligente prospere por todo el universo durante miles de millones de años, ¿cómo podemos estar seguros de que esta vida es consciente y capaz de apreciar lo que está ocurriendo? Si no lo fuese, se trataría, en palabras del famoso físico Erwin Schrödinger, de una «obra interpretada ante un auditorio de asientos vacíos, que no existiría para nadie, y que, por lo tanto, podría perfectamente decirse que no existe»?[100] Dicho de otro modo, si hacemos posible la existencia de descendientes tecnológicamente avanzados que creemos de forma errónea que son conscientes, ¿sería esto el apocalipsis zombi por antonomasia, que transformaría nuestra gran herencia cósmica en un mero desperdicio de espacio astronómico?

¿QUÉ ES LA CONSCIENCIA?

Muchas discusiones en torno a la consciencia generan más ruido que razones porque los interlocutores hablan sin escucharse realmente, y no son conscientes de que están usando definiciones distintas de lo que es la

consciencia. Como sucede con «vida» o «inteligencia», no existe una sola definición correcta e indiscutible de la palabra «consciencia», sino que existen muchas definiciones alternativas, con conceptos como sensibilidad, vigilia, autoconciencia, acceso a la información sensorial y capacidad de combinar información en un relato.[101] En nuestra exploración del futuro de la inteligencia, queremos tener una perspectiva lo más amplia e incluyente posible que no se limite a los tipos de consciencia biológica ya existentes. Por ese motivo, la definición que di en el capítulo 1, y a la que me estoy ajustando a lo largo del libro, es muy amplia:

consciencia = *experiencia subjetiva*

En otras palabras, si siente algo al ser usted ahora mismo, entonces tiene consciencia. Es esta particular definición de la consciencia la que llega al meollo de todas las preguntas relativas a la IA de la sección anterior: ¿se siente algo al ser Prometeo, AlphaGo o un coche Tesla autónomo?

Para que nos hagamos una idea de lo amplia que es nuestra definición de consciencia, fijémonos en que no menciona el comportamiento, la percepción, la autoconciencia, las emociones o la atención. Así pues, según esta definición también somos conscientes mientras soñamos, aunque no estemos despiertos ni tengamos acceso a la información sensorial y tampoco estemos sonámbulos haciendo cosas (¡o eso espero!). Análogamente, cualquier sistema que experimente dolor es consciente en este sentido, incluso aunque no pueda moverse. Nuestra definición deja abierta la posibilidad de que en el futuro puedan existir sistemas de IA conscientes, incluso aunque sea solo en forma de software y sin estar conectados a sensores o cuerpos robóticos.

Con esta definición, es difícil que no nos importe la conciencia. En palabras de Yuval Noah Harari en su libro *Homo Deus*:[102] «Si algún científico quiere argumentar que las experiencias subjetivas son irrelevantes, su problema es explicar por qué la tortura o la violación están mal sin hacer referencia a ninguna experiencia subjetiva». Sin esa referencia, todo es tan solo un montón de partículas elementales moviéndose de acá para allá de acuerdo con las leyes de la física. ¿Qué problema habría con eso?

¿Qué es exactamente lo que no entendemos sobre la consciencia? Pocos han reflexionado de manera tan profunda sobre esta cuestión como David Chalmers, un famoso filósofo australiano al que rara vez se le ve sin una sonrisa juguetona y una cazadora de cuero negra (que a mi mujer le gustó tanto que me regaló una parecida por Navidad). Se dejó llevar por el corazón y optó por la filosofía, a pesar de haber llegado a la final de la Olimpiada Internacional de Matemáticas (y a pesar de que el único notable que sacó en la universidad, la única mácula en un expediente repleto de sobresalientes, fue en una clase de introducción a la filosofía). De hecho, Chalmers parece del todo indiferente a los reproches o a las polémicas, y me asombra su capacidad de escuchar cortésmente las críticas desinformadas y equivocadas sobre su propio trabajo sin sentir siquiera la necesidad de responder.

Como ha destacado David, en relación con la mente hay en realidad dos misterios distintos. En primer lugar, está el misterio de cómo un cerebro procesa información, lo que David llama los problemas «fáciles». Por ejemplo, ¿cómo recibe, interpreta y responde el cerebro a la información sensorial? ¿Cómo puede dar cuenta de su estado interno mediante el lenguaje? Aunque estas cuestiones son de hecho sumamente difíciles, de acuerdo con nuestra definición no constituyen misterios de la consciencia, sino misterios de la inteligencia: plantean cómo recuerda, computa y aprende el cerebro. Es más, en la primera parte del libro vimos cómo los investigadores en IA han empezado a hacer avances importantes hacia la resolución de muchos de estos «problemas fáciles» con máquinas (desde jugar al go hasta conducir coches, analizar imágenes y procesar el lenguaje natural).

Por otra parte, está también el misterio de por qué tenemos una experiencia subjetiva que David denomina el problema *difícil*. Cuando conducimos, experimentamos colores, sonidos, emociones y una sensación de ser. Pero ¿por qué experimentamos algo en lugar de nada? ¿Experimenta algo un coche autónomo? Si competimos contra un coche autónomo, ambos recibimos información procedente de sensores, la procesamos y enviamos órdenes de movimiento. Pero experimentar la conducción subjetivamente es algo distinto desde un punto de vista lógico; ¿es opcional? y, en ese caso, ¿qué es lo que lo provoca?

Abordo este problema difícil de la consciencia desde un punto de vista físico. Desde esta perspectiva, una persona consciente no es más que comida reorganizada. ¿Por qué una forma de organización es consciente pero la otra no? Es más, la física nos enseña que la comida es simplemente una enorme cantidad de quarks y electrones organizados de una determinada manera. De modo que ¿qué disposiciones de partículas son conscientes y cuáles no?(30)

Lo que me gusta de esta perspectiva física es que transforma el problema difícil que los humanos llevamos milenios tratando de resolver en una versión más centrada que es más fácil abordar con los métodos de la ciencia. En lugar de empezar por el *problema* difícil de por qué una disposición de partículas puede sentirse consciente, hagámoslo con el *hecho* concreto de que algunas disposiciones de partículas se sienten conscientes y otras no. Por ejemplo, sabemos que las partículas que forman nuestro cerebro se encuentran en una disposición consciente en este momento, pero no cuando dormimos profundamente y no soñamos.

Esta perspectiva física conduce a tres preguntas difíciles y distintas sobre la consciencia, como se puede ver en la figura 8.1. En primer lugar, ¿qué propiedades de las disposiciones de partículas tienen consecuencias relevantes? Específicamente, ¿qué propiedades físicas distinguen los sistemas conscientes de los inconscientes? Si podemos responder a esto, podemos determinar qué sistemas de IA son conscientes. En el futuro más inmediato, también puede ayudar a los médicos de urgencias a determinar cuáles de los pacientes que no reaccionan son conscientes.

En segundo lugar, ¿de qué manera determinan las propiedades físicas cómo es la experiencia? En particular, ¿qué determina los *qualia*, los elementos básicos de la consciencia, como la rojez de una rosa, el sonido de un platillo, el olor de un filete, el sabor de una mandarina o el dolor de un pinchazo?(31)

Tercero, ¿por qué algo es consciente? En otras palabras, ¿hay alguna explicación, aún por descubrir, de por qué los pedazos de materia pueden ser conscientes, o se trata simplemente de un hecho inexplicable de la manera en que funciona el mundo?

En línea con David Chalmers, el informático Scott Aaronson, antiguo colega mío en el MIT, se ha referido alegremente a la primera pregunta como el «problema bastante difícil» (PBD). En esa misma línea, a los otros dos los llamaremos «problema aún más difícil» (PMD) y «problema realmente

difícil» (PRD), tal y como se ilustra en la figura 8.1.

¿ESTÁ LA CONSCIENCIA MÁS ALLÁ DEL ALCANCE DE LA CIENCIA?

Cuando la gente me dice que la investigación sobre la consciencia es un derroche inútil de tiempo, el principal argumento que me dan es que «no es científica» y nunca lo será. Pero ¿de verdad es así? El influyente filósofo austro-británico Karl Popper popularizó la máxima, ahora ampliamente aceptada, de que «si no es falsable, no es científico». Dicho de otro modo, lo que la ciencia pretende es poner a prueba las teorías contra las observaciones: si una teoría no se puede comprobar, ni siquiera en principio, entonces es lógicamente imposible refutarla alguna vez, lo que, de acuerdo con la frase de Popper, significa que no es científica.

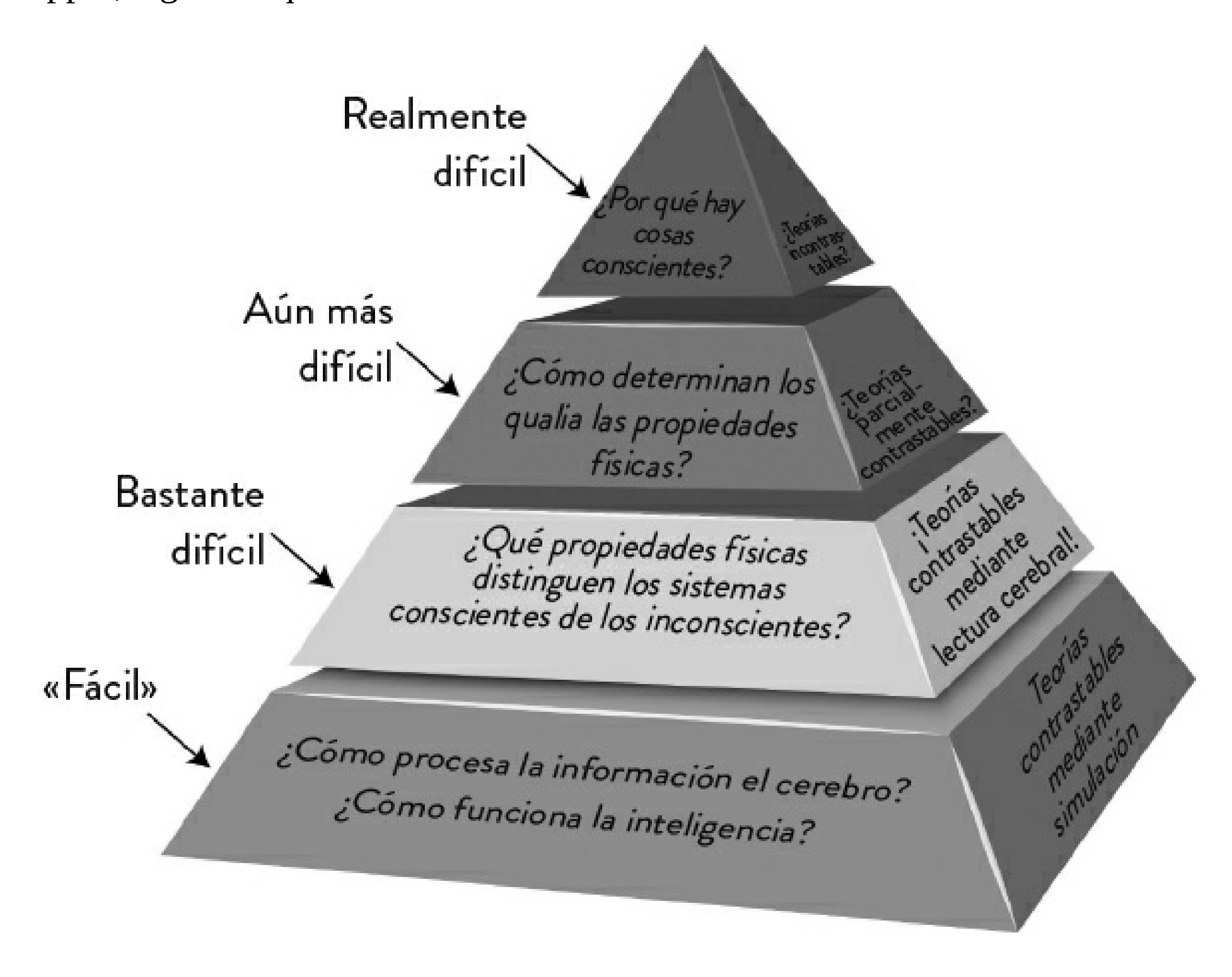


FIGURA 8.1. Comprender la mente implica una jerarquía de problemas. Lo que David Chalmers llama problemas «fáciles» puede plantearse sin mencionar la experiencia subjetiva. El hecho evidente de que algunos sistemas físicos, aunque no todos, sean conscientes plantea tres preguntas distintas. Si disponemos de una teoría para responder a la pregunta que define el «problema bastante difícil», entonces se puede comprobar experimentalmente. Si esta teoría funciona, podemos basarnos en ella para abordar las preguntas más difíciles.

¿Podría existir una teoría científica que dé respuesta a alguna de las tres preguntas sobre la consciencia recogidas en la figura 8.1? Permítame, por favor, que intente convencerlo de que la respuesta es un rotundo ¡SÍ!, al menos en el caso del problema bastante difícil: «¿Qué propiedades físicas distinguen los sistemas conscientes de los que no lo son?». Supongamos que alguien tiene una teoría tal que, dado cualquier sistema físico, responde a la pregunta de si el sistema es consciente con «sí», «no», «no es seguro». Conectamos nuestro cerebro a un dispositivo que mide el procesamiento de información en distintas regiones del mismo, e introducimos esa información en un programa de ordenador que utiliza la teoría de la consciencia para predecir qué partes de esa información son conscientes y nos muestra sus predicciones en tiempo real en un monitor, como en la figura 8.2. Primero, pensamos en una manzana. El monitor nos informa de que en nuestro cerebro hay información sobre una manzana de la que somos conscientes, pero también hay información en el tronco encefálico sobre nuestro pulso de la cual no tenemos conciencia. ¿Nos impresionaría saberlo? Aunque las dos primeras predicciones de la teoría eran correctas, decidimos llevar a cabo pruebas más rigurosas. Pensamos en nuestra madre y el ordenador nos informa de que en nuestro cerebro hay información sobre ella, pero que no somos conscientes de que dicha información exista. La teoría ha hecho una predicción errónea, lo que significa que hay que descartarla y tirarla a la papelera de historia de la ciencia, junto con la mecánica aristotélica, el éter lumínico, la cosmología geocéntrica e innumerables otras ideas fallidas. Esta es la cuestión clave: aunque la teoría era errónea, era científica. Si no lo hubiera sido, no habríamos podido ponerla a prueba y desecharla.

Habría quien podría criticar esta conclusión y decir que *ellos* no tienen evidencia de qué es aquello de lo que somos conscientes, o incluso de que seamos conscientes en absoluto: aunque nos oyen decir que somos conscientes, un zombi que no lo fuese podría decir exactamente lo mismo. Pero esto no hace que la teoría de la consciencia no sea científica, porque

podrían ponerse en nuestro lugar y comprobar si esta predice de forma correcta sus propias experiencias conscientes.

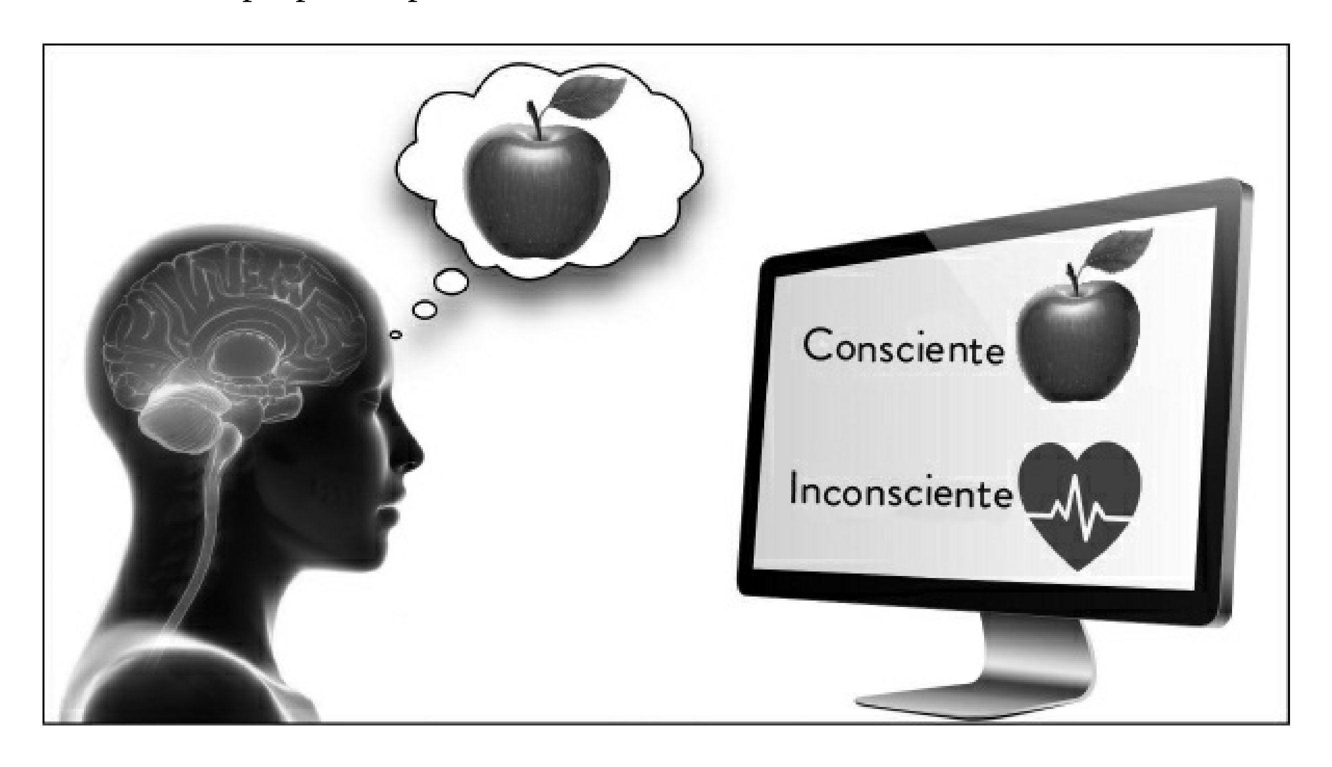


FIGURA 8.2. Supongamos que un ordenador mide la información que se procesa en nuestro cerebro y predice de qué regiones del mismo somos conscientes de acuerdo con una teoría de la consciencia. Podemos poner a prueba científicamente esta teoría comprobando si sus predicciones son correctas y coinciden con nuestra experiencia subjetiva.

Por otro lado, si la teoría se niega a hacer cualquier predicción, y responde simplemente «no es seguro» cada vez que se le pregunta, entonces no puede ponerse a prueba, y por lo tanto no es científica. Esto podría suceder porque la teoría fuese válida solo en algunas situaciones, porque los cálculos requeridos son demasiado difíciles para llevarlos a cabo en la práctica o porque los sensores cerebrales no funcionan bien. Las teorías científicas más conocidas actualmente tienden a situarse en algún punto intermedio, y proporcionan respuestas comprobables a algunas de nuestras preguntas, pero no a todas. Por ejemplo, la teoría fundamental de la física se negará a dar respuesta a preguntas sobre sistemas que sean a la vez extremadamente pequeños (que requieren de la mecánica cuántica) y extremadamente pesados (para los cuales se necesita la relatividad general), porque todavía no hemos descubierto qué ecuaciones matemáticas usar en este caso. Esta teoría fundamental también se negará a predecir las masas exactas de todos los

átomos posibles; en este caso, creemos que tenemos las ecuaciones necesarias, pero no hemos logrado calcular con precisión sus soluciones. Cuanto más peligrosamente vive una teoría, al arriesgarse a hacer predicciones comprobables, más útil es, y más peso le damos si sobrevive a todos nuestros intentos de acabar con ella. Sí, solo podemos comprobar algunas de las predicciones de las teorías de la consciencia, pero lo mismo sucede con todas las teorías físicas. Así pues, no perdamos el tiempo quejándonos de lo que no podemos poner a prueba, y dediquémoslo a comprobar lo que sí podemos comprobar.

En resumen, cualquier teoría que prediga qué sistemas físicos son conscientes (el problema bastante difícil) es científica siempre que pueda predecir cuáles de nuestros procesos cerebrales son conscientes. Sin embargo, la cuestión de la comprobabilidad es menos clara para las cuestiones que aparecen en la figura 8.1. ¿Qué significaría que una teoría predijese cómo experimentamos subjetivamente el color rojo? Y si una teoría pretendiese explicar por qué existe tal cosa como la consciencia, ¿cómo se comprobaría experimentalmente? El mero hecho de que estas preguntas sean difíciles no basta para que debamos eludirlas, por lo que volveremos sobre ellas más adelante. Pero, cuando tenemos ante nosotros varias cuestiones sin resolver relacionadas entre sí, considero prudente empezar a abordar en primer lugar la más sencilla. Por este motivo, mi investigación sobre la consciencia en el MIT se centra exclusivamente en la base de la pirámide de la figura 8.1. Hace poco, comenté esta estrategia con Piet Hut, físico en Princeton, que bromeando comentó que tratar de construir el vértice de la pirámide antes que la base sería como preocuparse por la interpretación de la mecánica cuántica antes de haber descubierto la ecuación de Schrödinger, la base matemática que nos permite predecir los resultados de nuestros experimentos.

Cuando se debate sobre qué es lo que está fuera del alcance de la ciencia, es importante recordar que la respuesta varía con el tiempo. Hace cuatro siglos, Galileo Galilei estaba tan deslumbrado por las teorías físicas de base matemática que describió la naturaleza como «un libro escrito en el lenguaje de las matemáticas». Si dejaba caer una uva y una avellana, podía predecir con precisión la forma de sus trayectorias y cuándo llegarían al suelo. Pero no tenía ni idea de por qué una era verde y la otra marrón, o por qué una era blanda y la otra dura. Estas características del mundo quedaban fuera del alcance de la ciencia de la época. ¡Pero no para siempre! Cuando James Clerk

Maxwell descubrió en 1861 las ecuaciones que llevan su nombre, resultó evidente que la luz y los colores también podían comprenderse matemáticamente. Ahora sabemos que la ya mencionada ecuación de Schrödinger, descubierta en 1925, puede usarse para predecir todas las propiedades de la materia, incluso qué cosas son blandas o duras. Si los avances teóricos han permitido hacer cada vez más predicciones científicas, el progreso tecnológico ha hecho posibles un número cada vez mayor de comprobaciones experimentales: casi todo lo que estudiamos hoy en día usando telescopios, microscopios o colisionadores de partículas estuvo en otra época más allá del alcance de la ciencia. Dicho de otra manera, el ámbito de la ciencia se ha ampliado espectacularmente desde la época de Galileo, y ha pasado de incluir una minúscula proporción de todos los fenómenos a un gran porcentaje de los mismos, incluidas las partículas subatómicas, los agujeros negros y nuestros orígenes cósmicos hace 13.800 millones de años. Esto suscita la siguiente pregunta: ¿qué más queda?

En mi opinión, la consciencia es el elefante en la habitación. No solo sabemos que somos conscientes, sino que es *lo único* que sabemos con total certeza; todo lo demás es inferencia, como señaló René Descartes en tiempos de Galileo. ¿Conseguirá en algún momento el progreso teórico y tecnológico arrastrar firmemente incluso la consciencia dentro del dominio de la ciencia? No lo sabemos, igual que Galileo no sabía si algún día entenderíamos la luz y la materia.(32) Solo hay una cosa segura: no lo conseguiremos si no lo intentamos. Esta es la razón por la cual tanto yo, como muchos otros científicos de todo el mundo, estamos poniendo todo nuestro empeño en formular y poner a prueba teorías de la consciencia.

INDICIOS EXPERIMENTALES SOBRE LA CONSCIENCIA

En este mismo instante tiene lugar en nuestras cabezas una ingente cantidad de procesamiento de información. ¿Qué parte de él es consciente y cuál no? Antes de abordar las teorías de la consciencia y lo que predicen, echemos un vistazo a lo que nos han enseñado hasta la fecha los experimentos, desde los tradicionales con tecnologías poco o nada avanzadas hasta las mediciones del cerebro realizadas con tecnología punta.

¿Qué comportamientos son conscientes?

Si multiplicamos mentalmente 32 por 17, somos conscientes de muchos de los entresijos del cálculo. Pero supongamos que, en lugar de eso, nos muestran un retrato de Albert Einstein y nos piden que digamos el nombre de la persona en cuestión. Como vimos en el capítulo 2, esta es también una tarea computacional: nuestro cerebro está evaluando una función, cuyos datos de entrada los constituyen información procedente de nuestros ojos sobre una gran cantidad de píxeles de colores y cuya salida es información dirigida a los músculos que controlan nuestra boca y nuestras cuerdas vocales. Para los informáticos, se trata de «clasificación de imágenes» seguida de «síntesis de habla». Aunque esta computación es muchísimo más complicada que la multiplicación anterior, podemos hacerla más rápido, sin aparente esfuerzo, y sin ser conscientes de los detalles de *cómo* la hacemos. Nuestra experiencia subjetiva consiste solamente en mirar la imagen, experimentar una sensación de reconocimiento y oírnos decir «Einstein».

Los psicólogos saben desde hace tiempo que también podemos realizar de manera inconsciente una amplia variedad de otras tareas y comportamientos, desde pestañeos reflejos hasta respirar, alcanzar algo y agarrarlo, o mantener el equilibrio. Normalmente, tenemos conocimiento consciente de lo que hemos hecho, pero no de cómo lo hicimos. Por otra parte, comportamientos en los que intervienen una situación poco habitual, autocontrol, reglas lógicas complicadas, razonamiento abstracto o manipulación del lenguaje suelen ser conscientes. Se los conoce como *correlatos conductuales de la consciencia*, y están estrechamente vinculados a la forma de pensar lenta, controlada y trabajosa que los psicólogos llaman «sistema 2».[103]

También se sabe que, mediante la práctica intensiva, muchas rutinas se pueden convertir de conscientes en inconscientes; por ejemplo, caminar, nadar, montar en bicicleta, conducir, escribir, afeitarse, atarse los zapatos, jugar a videojuegos y tocar el piano.[104] De hecho, es bien sabido que los expertos ejercen sus especialidades mejor cuando se encuentran en un estado de «flujo», y son conscientes únicamente de lo que está sucediendo a un nivel más elevado, pero no de los detalles de bajo nivel de cómo lo están haciendo. Por ejemplo, intente leer la siguiente oración siendo consciente de cada letra, como cuando aprendió a leer. ¿Puede sentir cómo es mucho más lento, en

comparación con cuando somos simplemente conscientes del texto al nivel de las palabras o las ideas?

De hecho, no solo parece que es posible el procesamiento inconsciente de información, sino que es más la regla que la excepción. Hay evidencias que sugieren que de los aproximadamente 10^7 bits de información que llegan a nuestro cerebro cada segundo procedentes de nuestros órganos sensoriales, solo podemos ser conscientes de una minúscula proporción, que se estima entre 10 y 50 bits.[105] Esto parece indicar que el procesamiento de información del que tenemos conocimiento consciente no es más que la punta del iceberg.

Tomados en su conjunto, estos indicios han llevado a algunos investigadores a proponer que el procesamiento de información consciente debe considerarse como el presidente ejecutivo de nuestra mente, que solo se ocupa de las decisiones más importantes, que requieren análisis complejos de datos de todo el cerebro.[106] Esto explicaría por qué, como el presidente ejecutivo de una compañía, por lo general no quiere entretenerse en saber todo lo que están haciendo sus subordinados, pero podría averiguarlo si quisiese. Para experimentar esta atención selectiva en acción, detengámonos en esa última palabra, «quisiese»: fijemos la mirada en el punto sobre la primera «i» y, sin mover los ojos, desplacemos la atención del punto a la letra entera, y luego a la palabra completa. Aunque la información en nuestra retina siguió siendo la misma, nuestra experiencia consciente cambió. La metáfora del presidente ejecutivo también explica por qué la experiencia se hace inconsciente: después de aprender esforzadamente a leer y escribir, el presidente ejecutivo delega estas tareas rutinarias en subordinados inconscientes, para así poder centrarse en nuevos problemas de alto nivel.

¿Dónde ocurre la consciencia?

Ingeniosos experimentos y análisis sugieren que la consciencia está limitada no solo a ciertos comportamientos, sino también a ciertas regiones del cerebro. ¿Cuáles son los principales sospechosos? Muchos de los primeros indicios se obtuvieron de pacientes con lesiones cerebrales: daños cerebrales localizados provocados por accidentes, derrames cerebrales, tumores o

infecciones. Pero a menudo no permitían sacar conclusiones. Por ejemplo, el hecho de que las lesiones en la parte posterior del cerebro puedan causar ceguera ¿significa que este es el lugar donde radica la conciencia visual, o simplemente significa que la información visual pasa por allí en su tránsito hacia donde posteriormente se hará consciente, igual que antes ha pasado por los ojos?

Aunque las lesiones y las intervenciones quirúrgicas no han identificado las ubicaciones exactas de las experiencias conscientes, sí han ayudado a reducir las opciones. Por ejemplo, sé que, aunque experimento dolor en la mano como si en realidad se produjese en ella, la experiencia del dolor debe ocurrir en otro lugar, porque un cirujano eliminó el dolor de mi mano sin hacer nada con ella: se limitó a anestesiar los nervios de mi hombro. Además, algunos mancos experimentan un dolor fantasma que sienten como si se produjera en su mano inexistente. Otro ejemplo: una vez me di cuenta de que, si miraba solo con el ojo derecho, desaparecía parte de mi campo visual; un médico determinó que la retina se me estaba desprendiendo y la volvió a fijar. Por su parte, los pacientes con ciertas lesiones cerebrales experimentan *heminegligencia*, que hace que pierdan información de la mitad de su campo visual, pero ni siquiera son conscientes de ello; por ejemplo, ni ven ni se comen la comida de la mitad izquierda de su plato. Es como si la consciencia de la mitad de su mundo hubiera desaparecido. Pero ¿se supone que esas regiones del cerebro dañadas generan la experiencia espacial, o simplemente transmiten la información espacial a los lugares donde radica la consciencia, como sucedía con mi retina?

El pionero neurocirujano estadounidense-canadiense Wilder Penfield descubrió en la década de 1930 que sus pacientes sentían que les tocaban distintas partes del cuerpo cuando les estimulaban eléctricamente determinadas regiones del cerebro pertenecientes a lo que ahora se conoce como corteza somatosensorial (figura 8.3).[107] Penfield también descubrió que sus pacientes movían involuntariamente diferentes partes de su cuerpo cuando estimulaba regiones del cerebro situadas en lo que ahora se llama corteza motora. Pero ¿significa eso que el procesamiento de la información en estas áreas del cerebro corresponde a la consciencia del tacto y del movimiento?

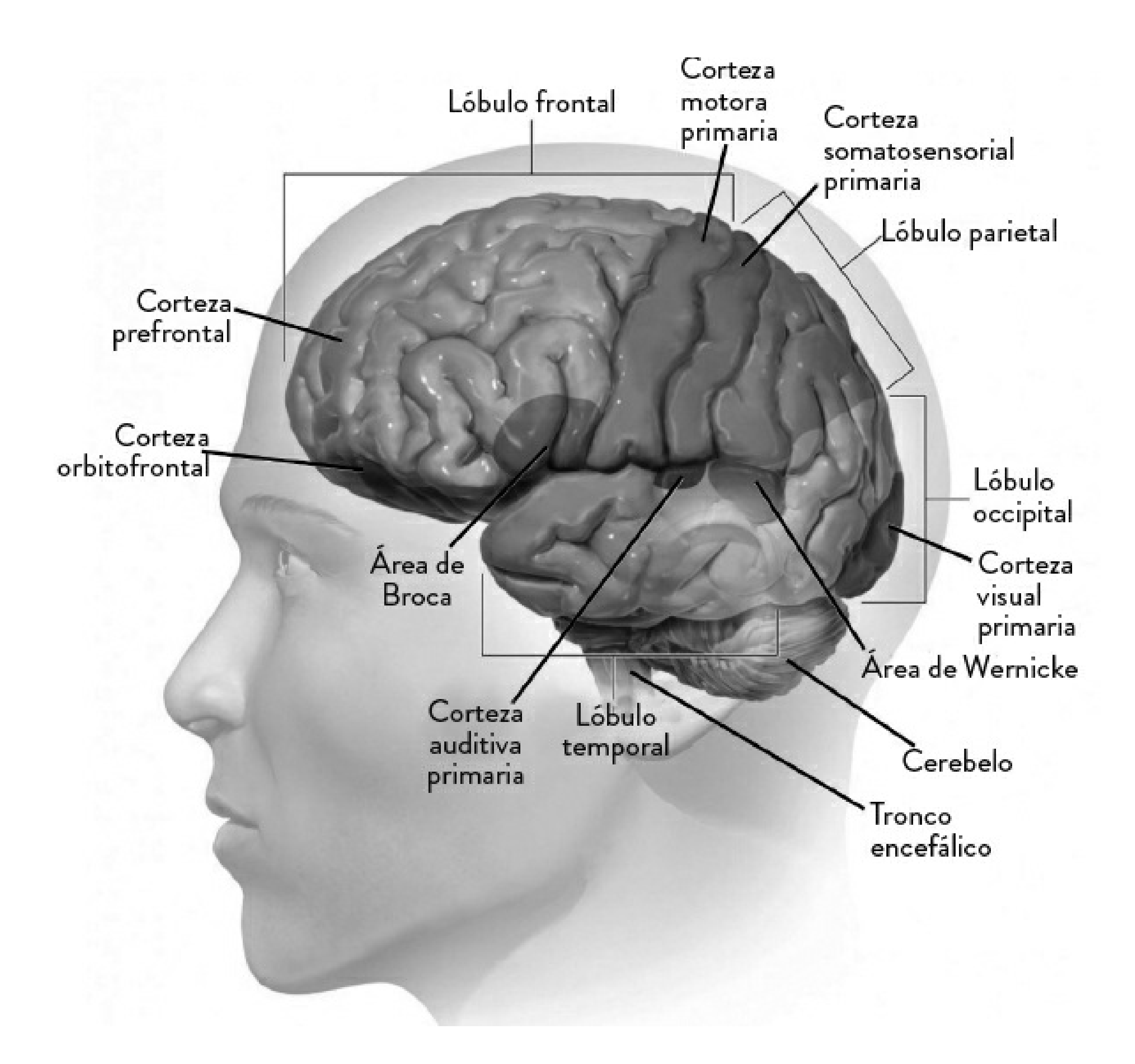


FIGURA 8.3. Las cortezas visual, auditiva, somatosensorial y motora están relacionadas con la vista, el oído, el sentido del tacto y la activación del movimiento, respectivamente, pero eso no prueba que sea en ellas donde ocurre la *consciencia* de la vista, del oído, del tacto y del movimiento. De hecho, investigaciones recientes sugieren que la corteza visual primaria es por completo inconsciente, igual que el cerebelo y el tronco encefálico. Imagen cortesía de Lachina (<www.lachina.com>).

Por fortuna, la tecnología moderna nos ofrece indicios mucho más detallados. Aunque aún no estamos ni remotamente cerca de ser capaces de medir cada activación de las aproximadamente cien mil millones de neuronas que hay en nuestro cerebro, la tecnología de lectura cerebral está avanzando a marchas forzadas, con técnicas de nombres tan imponentes como IRMf, EEG, MEG, ECoG, electrofisiología y detección de voltaje fluorescente. La

IRMf (siglas de Imagen por Resonancia Magnética funcional) mide las propiedades magnéticas de los núcleos de hidrógeno para crear un mapa tridimensional del cerebro más o menos cada segundo, con una resolución del orden de un milímetro. La EEG (electroencefalografía) y la MEG (magnetoencefalografía) miden el campo eléctrico y magnético en el exterior de la cabeza para crear un mapa del cerebro mil veces por segundo, pero con baja resolución, incapaz de distinguir detalles de menos de unos pocos centímetros. Si es usted aprensivo, agradecerá que estas tres técnicas sean no invasivas. Si no le importa que le abran el cráneo, dispone de más opciones. La ECoG (electrocorticografía) consiste en colocar un centenar de cables en la superficie del cerebro, mientras que en la electrofisiología se insertan en el cerebro microcables, a veces más finos que un cabello humano, para registrar voltajes de hasta mil puntos a la vez. Muchos pacientes epilépticos pasan días en el hospital, mientras se les aplica ECoG para descubrir qué parte de su cerebro está provocando las convulsiones y debe ser resecada, y amablemente acceden a que, durante ese tiempo, los neurocientíficos realicen sobre ellos experimentos relacionados con la consciencia. Por último, la detección de voltaje fluorescente implica la manipulación genética de las neuronas para que emitan destellos de luz al activarse, lo que permite medir su actividad con un microscopio. De todas las técnicas, esta es la que permite monitorizar rápidamente la mayor cantidad de neuronas, al menos en animales con cerebros transparentes, como el gusano *C. elegans*, con sus 302 neuronas, y las larvas de pez cebra, con sus alrededor de 100.000.

Aunque Francis Crick previno a Christof Koch en contra de estudiar la consciencia, este se negó a darse por vencido y acabó haciendo cambiar de opinión a Francis. En 1990, escribieron un artículo seminal sobre lo que denominaron «correlatos neuronales de la consciencia» (CNC), en el que se preguntaban qué procesos cerebrales concretos correspondían a experiencias conscientes. Durante miles de años, los pensadores únicamente habían tenido acceso al procesamiento de información que tenía lugar en sus cerebros a través de su experiencia y su comportamiento subjetivos. Crick y Koch señalaron que la tecnología de lectura cerebral estaba proporcionando de repente acceso independiente a esta información, permitiendo así el estudio científico de qué procesamiento de información correspondía a qué experiencia consciente. Como cabía esperar, las mediciones que la tecnología ha posibilitado han hecho que la búsqueda de CNC haya pasado a ser una

parte bastante común de la neurociencia, una cuyas miles de publicaciones llegan incluso a las revistas más prestigiosas.[108]

¿Qué conclusiones se han alcanzado hasta ahora? Para hacernos una idea de cómo es el trabajo detectivesco de la búsqueda de CNC, empecemos por plantearnos si nuestra retina es consciente, o si no es más que un sistema zombi que registra información visual, la procesa y la envía a otro sistema posterior en el cerebro donde tiene lugar nuestra experiencia consciente. En la imagen de la izquierda de la figura 8.4, ¿qué cuadrado es más oscuro: el marcado como A o el B? El A, ¿verdad? Pues no, lo cierto es que ambos son del mismo color, lo cual puede comprobarse mirándolos a través de un pequeño resquicio entre nuestros dedos. Esto demuestra que nuestra experiencia visual no puede residir por completo en la retina, ya que, si así fuera, los cuadrados se verían iguales.

Ahora fijémonos en la imagen de la derecha de la figura 8.4. ¿Vemos dos mujeres o una vasija? Si miramos la imagen durante un rato, experimentaremos subjetivamente ambas posibilidades de forma sucesiva, aunque la información que llega a nuestra retina es siempre la misma. Si se mide lo que sucede en nuestro cerebro durante ambas situaciones, es posible determinar qué las diferencia; y no es la retina, que se comporta idénticamente en ambos casos.

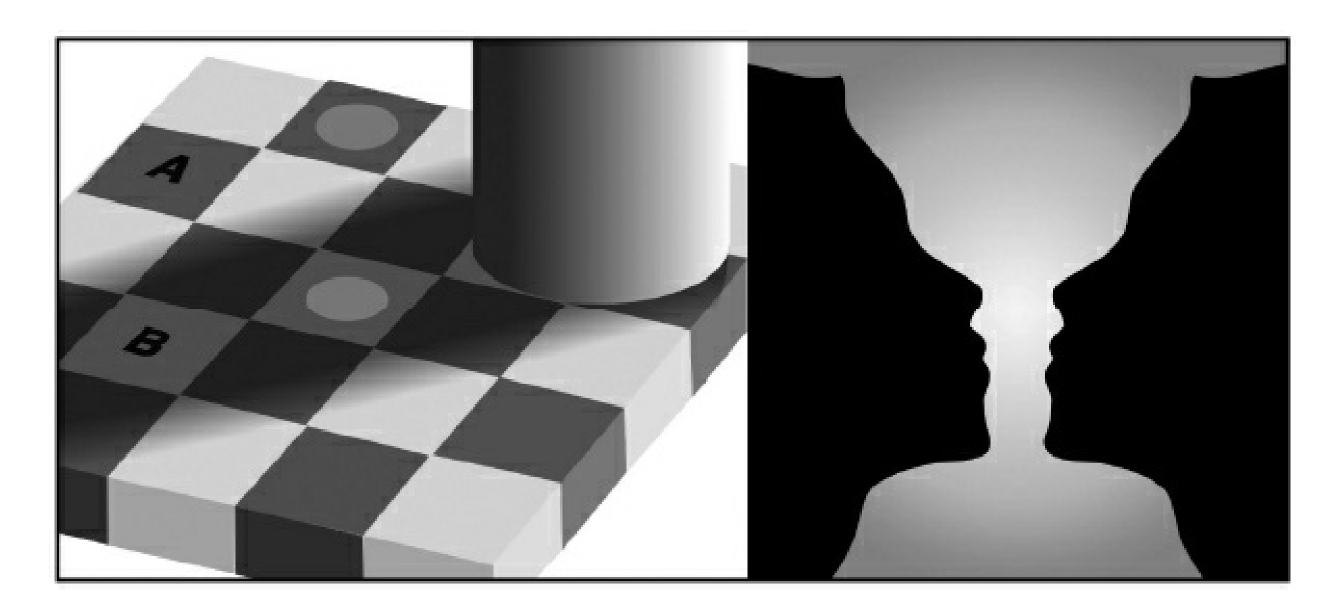


FIGURA 8.4. ¿Qué cuadrado es más oscuro, el A o el B? ¿Qué vemos a la derecha, una vasija, dos mujeres, o ambas imágenes sucesivamente? Ilusiones como estas demuestran que la consciencia visual no puede estar en los ojos o en alguna otra de las primeras etapas del sistema visual, porque no depende

tan solo de lo que hay en la imagen.

El golpe de gracia para la hipótesis de la retina consciente proviene de una técnica llamada «supresión continua con flash», propuesta por Christof Koch, Stanislas Dehaene y sus colaboradores: se ha descubierto que, si con un solo ojo vemos una secuencia complicada de patrones que cambian rápidamente, esto distrae a nuestro sistema visual hasta tal punto que, si al otro ojo se le muestra una imagen estática, no tendremos ninguna consciencia de ella.[109] En resumen, podemos tener una imagen visual en la retina sin experimentarla, y podemos (mientras soñamos) experimentar una imagen sin que esté en la retina. Esto demuestra que las retinas no albergan la consciencia visual más de lo que lo hace una cámara de vídeo, aunque sí realizan complicados cálculos en los que intervienen más de cien millones de neuronas.

Los investigadores en CNC también utilizan la supresión continua con flash, ilusiones visuales/auditivas inestables y otros trucos para localizar qué regiones del cerebro son responsables de cada una de nuestras experiencias conscientes. La estrategia básica consiste en comparar lo que hacen las neuronas en dos situaciones en las que básicamente todo es igual (incluida la información sensorial), salvo la experiencia consciente. Las partes del cerebro en las que se obtienen mediciones distintas se identifican como CNC.

Estas investigaciones en CNC han demostrado que ninguna parte de la consciencia reside en el intestino, aunque es ahí donde está situado el sistema nervioso entérico, con sus nada menos que 500 millones de neuronas que calculan cuál es la manera óptima de digerir los alimentos; sensaciones como las de hambre o náuseas no se producen ahí, sino en el cerebro. Análogamente, parece que ninguna parte de la consciencia reside tampoco en el tronco encefálico, la parte inferior del cerebro que conecta con la médula espinal y controla la respiración, el ritmo cardiaco y la presión sanguínea. Más sorprendente resulta que la consciencia tampoco parezca extenderse al cerebelo (figura 8.3), que contiene en torno a dos tercios de todas nuestras neuronas: los pacientes cuyo cerebelo resulta destruido experimentan dificultades en el habla y torpeza de movimientos similares a las provocadas por la embriaguez, pero no dejan de ser plenamente conscientes.

La pregunta de qué partes del cerebro son responsables de la consciencia sigue estando abierta y es objeto de polémica. Un estudio reciente sobre CNC

sugiere que la consciencia reside principalmente en una «zona caliente», de la que forman parte el tálamo (cerca del centro del cerebro) y la parte posterior de la corteza (la capa externa del cerebro, que consiste en una lámina arrugada de seis capas que, si se extendiese, tendría el tamaño de una servilleta grande).[110] Ese mismo estudio sugiere que la corteza visual primaria en la parte posterior de la cabeza es una excepción a lo anterior, y es tan inconsciente como los globos oculares y las retinas, lo cual ha dado pie a la polémica.

¿Cuándo ocurre la consciencia?

Hasta ahora, nos hemos fijado en los indicios experimentales relativos a qué tipos de procesamiento de información son conscientes y dónde ocurre la consciencia. Pero *¿cuándo* ocurre? Cuando yo era niño, solíamos pensar que tomamos consciencia de los acontecimientos a medida que suceden, sin ningún desfase o retardo temporal. Aunque sigo sintiendo que así es como sucede, es evidente que no puede ser cierto, porque mi cerebro tarda un tiempo en procesar la información que recibe de los órganos sensoriales. Los investigadores en CNC han medido cuánto tiempo, y la conclusión de Christof Koch es que transcurre aproximadamente un cuarto de segundo desde que la luz procedente de un objeto complejo entra en el ojo hasta que lo percibimos de forma consciente, viéndolo como lo que es.[111] Esto significa que, si conducimos por la autopista a 90 kilómetros por hora y de pronto vemos una ardilla unos pocos metros por delante, ya es demasiado tarde para hacer algo, porque ya la hemos atropellado.

Resumiendo, nuestra consciencia vive en el pasado; Christof Koch estima que lleva un retardo de un cuarto de segundo respecto al mundo exterior. Curiosamente, muchas veces podemos reaccionar ante las cosas más rápido de lo que somos conscientes de ellas, lo que demuestra que el procesamiento de información encargado de nuestras reacciones más rápidas debe ser inconsciente. Por ejemplo, cuando un objeto extraño se aproxima a nuestro ojo, un pestañeo reflejo puede cerrar el párpado en tan solo una décima de segundo. Es como si uno de los sistemas del cerebro recibiese la información inquietante procedente del sistema visual, computase que el ojo está en peligro de recibir un impacto, enviase un mensaje a los músculos del ojo para

que parpadeasen y, simultáneamente, se comunicase con la parte consciente del cerebro para decirle: «Eh, vamos a parpadear». Para cuando este mensaje ha sido leído e incluido en nuestra experiencia consciente, el parpadeo ya se ha producido.

De hecho, el sistema que lee ese mensaje es bombardeado constantemente con comunicaciones procedentes de todo el cuerpo, algunas con más retardo que otras. Las señales nerviosas tardan más en llegar al cerebro desde los dedos de la mano que desde la cara debido a la distancia, y tardamos más en analizar imágenes que sonidos porque lo primero es más complicado (esta es la razón por la que las competiciones olímpicas comienzan con un disparo en lugar de con una señal visual). Pero si nos tocamos la nariz, experimentamos la sensación en la nariz y en la yema de los dedos como simultánea, y, si damos una palmada, la vemos, oímos y sentimos exactamente en el mismo instante.[112] Esto significa que toda nuestra experiencia consciente de un evento no se crea hasta que los mensajes más retrasados se han recibido y analizado.

Una famosa familia de experimentos en CNC propuesta inicialmente por el fisiólogo Benjamin Libet ha demostrado que los tipos de acciones que podemos realizar inconscientemente no se limitan a respuestas rápidas como parpadeos y remates de ping-pong, sino que también incluyen ciertas decisiones que podríamos atribuir al libre albedrío. En ocasiones, las mediciones del cerebro pueden predecir nuestra decisión antes de que seamos conscientes de haberla realizado.[113]

TEORÍAS DE LA CONSCIENCIA

Acabamos de ver que, aunque aún no comprendemos la consciencia, disponemos de una cantidad asombrosa de datos experimentales sobre diversos aspectos de la misma. Pero todos estos datos proceden de *cerebros*; ¿cómo pueden enseñarnos algo sobre la consciencia en *máquinas*? Esto requiere una extrapolación más allá de nuestro actual dominio experimental; es decir, requiere una *teoría*.

¿Por qué una teoría?

Para entender por qué, comparemos las teorías de la consciencia con las teorías de la gravedad. Los científicos empezaron a tomarse en serio la teoría de la gravedad de Newton porque obtenían de la misma más de lo que ponían en ella: unas ecuaciones sencillas que cabían en una servilleta permitían predecir con precisión el resultado de cualquier experimento relacionado con la gravedad que se hubiese llevado a cabo jamás. Por lo tanto, también se tomaron en serio sus predicciones mucho más allá del dominio en el que se habían comprobado, y resultó que estas audaces extrapolaciones funcionaron incluso para los movimientos de galaxias en cúmulos de millones de años luz de diámetro. Sin embargo, las predicciones eran ligerísimamente incorrectas para la órbita de Mercurio alrededor del Sol. Los científicos empezaron a tomarse en serio la teoría mejorada de Einstein, la teoría de la relatividad general, porque era aún más elegante y escueta, y predecía correctamente lo que la de Newton no hacía. En consecuencia, también tomaron en serio sus predicciones más allá del dominio donde se habían comprobado, para fenómenos tan exóticos como los agujeros negros, las ondas gravitacionales en el propio tejido del espacio-tiempo y la expansión del universo desde sus ardientes orígenes, todos ellos posteriormente confirmados a través de experimentos.

En ese mismo sentido, si una teoría matemática de la consciencia, cuyas ecuaciones cupiesen en una servilleta, pudiera predecir con éxito los resultados de todos los experimentos que realizamos en el cerebro, comenzaríamos a tomarnos en serio no solo la teoría en sí, sino también sus predicciones para la consciencia más allá del cerebro; por ejemplo, en máquinas.

La consciencia desde un punto de vista físico

Aunque algunas teorías de la consciencia se remontan a la Antigüedad, la mayoría de las teorías modernas se basan en la neuropsicología y la neurociencia, e intentan explicar y predecir la consciencia en función de eventos neuronales que tienen lugar en el cerebro.[114] Aunque estas teorías han predicho con éxito algunos correlatos neuronales de la consciencia, no pueden hacer predicciones sobre la consciencia de las máquinas, ni aspiran a

ello. Para dar el salto de los cerebros a las máquinas, necesitamos generalizar de los CNC a los CPC (*correlatos físicos de la consciencia*), definidos como los patrones de partículas en movimiento que son conscientes. Porque, si una teoría puede predecir correctamente lo que es consciente y lo que no haciendo referencia tan solo a componentes físicos como las partículas elementales y los campos de fuerza, entonces puede hacer predicciones no solo para cerebros sino también para cualquier otra disposición de la materia, incluidos los futuros sistemas de IA. Así pues, adoptemos un punto de vista físico: ¿qué disposiciones de partículas son conscientes?

Pero, en realidad, esto plantea otra cuestión: ¿cómo puede algo tan complejo como la consciencia estar compuesto de algo tan simple como las partículas? Creo que es porque se trata de un fenómeno que tiene propiedades que van más allá de las de sus partículas. En física, estos fenómenos se denominan «emergentes».[115] Tratemos de entender esto fijándonos en un fenómeno emergente más simple que la consciencia: la humedad.

Una gota de agua es húmeda, pero un cristal de hielo y una nube de vapor no lo son, aunque están hechos de idénticas moléculas de agua. ¿Por qué? Porque la propiedad de la humedad depende únicamente de la disposición de las moléculas. No tiene ningún sentido afirmar que una sola molécula de agua es húmeda, porque el fenómeno de la humedad aparece solamente cuando hay muchas moléculas, organizadas según un patrón que llamamos «líquido». De manera que sólidos, líquidos y gases son todos ellos fenómenos emergentes: son más que la suma de sus partes, porque poseen propiedades que van más allá de las propiedades de sus partículas. Tienen propiedades de las que las partículas carecen.

Como sucede con sólidos, líquidos y gases, yo pienso que la consciencia es un fenómeno emergente, con propiedades que van más allá de las de sus partículas. Por ejemplo, caer en un sueño profundo extingue la consciencia, simplemente mediante una reorganización de las partículas. Del mismo modo, mi consciencia desaparecería si muriese congelado, lo que reorganizaría mis partículas de una manera más desafortunada.

Cuando juntamos un montón de partículas para crear algo, ya sea agua o un cerebro, aparecen nuevos fenómenos con propiedades observables. A los físicos nos encanta estudiar estas propiedades emergentes, que a menudo se pueden identificar mediante un reducido conjunto de números que podemos medir, magnitudes tales como lo viscosa que es una sustancia, lo compresible

que es, etcétera. Por ejemplo, si una sustancia es tan viscosa que es rígida, decimos que es sólida; de lo contrario, será fluida. Y si un fluido no es compresible, decimos que es un líquido; de lo contrario, lo llamamos gas o plasma, dependiendo de en qué medida conduzca la electricidad.

Consciencia como información

¿Podrían existir magnitudes análogas que cuantifiquen la consciencia? El neurocientífico italiano Giulio Tononi ha propuesto una de estas magnitudes, que él llama *información integrada* —y representa mediante la letra griega Φ (*fi*)—, que básicamente mide cuánto saben las distintas partes de un sistema las unas de las otras (véase la figura 8.5).

Conocí a Giulio en una conferencia de física que tuvo lugar en 2014 en Puerto Rico, y a la que invité tanto a él como a Christof Koch, y me pareció un consumado hombre del Renacimiento que habría podido codearse con Galileo y Leonardo de Vinci. Su actitud tranquila no ocultaba su extraordinario conocimiento del arte, la literatura y la filosofía, y lo precedía su reputación culinaria: un periodista de televisión cosmopolita me había contado recientemente que Giulio había preparado en pocos minutos la ensalada más deliciosa que había probado en su vida. Enseguida me di cuenta de que, tras de su actitud relajada, había un intelecto intrépido que se dejaría guiar por la evidencia hasta donde fuera necesario, independientemente de las ideas preconcebidas y los tabúes dominantes. Así como Galileo había mantenido su teoría matemática del movimiento a pesar de la presión de las autoridades para que no pusiese en duda el geocentrismo, Giulio había desarrollado la teoría de la consciencia matemáticamente más precisa hasta la fecha, la *teoría de la información integrada* (TII).

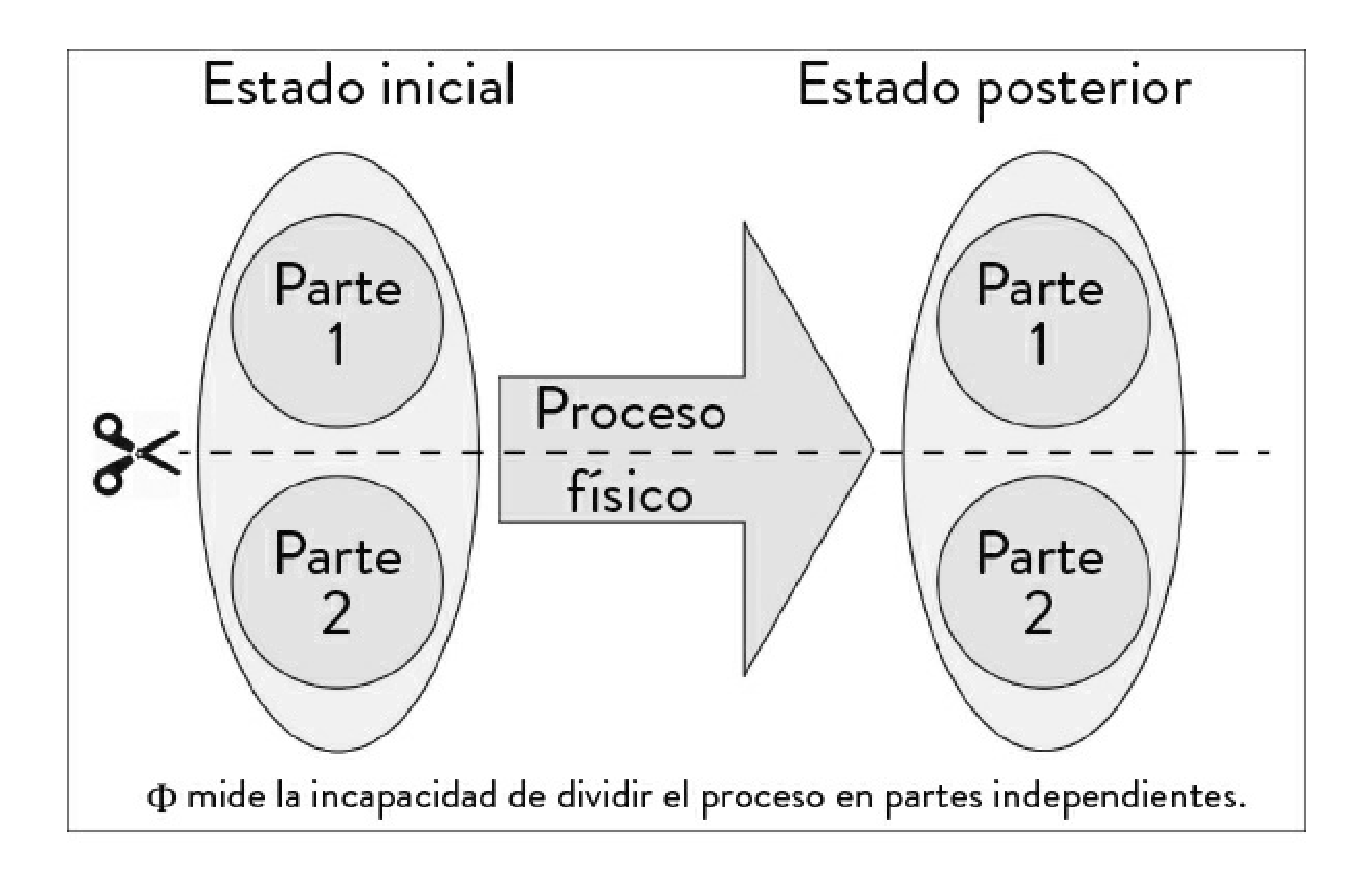


FIGURA 8.5. Dado un proceso físico que, con el paso del tiempo, transforma el estado inicial de un sistema en un nuevo estado, su información integrada Φ mide la incapacidad de dividir el proceso en partes independientes. Si el estado futuro de cada parte depende solo de su propio pasado, no de lo que la otra parte ha estado haciendo, entonces Φ = 0, y lo que estamos considerando un sistema son en realidad dos sistemas independientes que no se comunican entre sí.

Yo llevaba décadas sosteniendo que la consciencia es cómo se siente la información cuando se procesa de determinadas maneras complejas.[116] La TII coincide con mi afirmación, pero sustituye mi imprecisa frase de «determinadas maneras complejas» por una definición precisa: el procesamiento de información debe ser integrado; esto es, Φ debe ser grande. El razonamiento de Giulio para afirmar lo anterior es tan potente como sencillo: el sistema consciente necesita estar integrado en un todo unificado, porque si, por el contrario, estuviese formado por dos partes independientes, entonces se sentirían como dos entidades conscientes distintas en lugar de solo una. Dicho de otro modo, si una parte consciente de un cerebro u ordenador no puede comunicarse con el resto, el resto no puede formar parte de su experiencia subjetiva.

Giulio y sus colaboradores han medido una versión simplificada de Φ

usando EEG para medir la respuesta del cerebro a la estimulación magnética. Su «detector de consciencia» funciona realmente bien: determinó que los pacientes estaban conscientes cuando estaban despiertos o soñando, pero inconscientes cuando estaban anestesiados o en un sueño profundo. Incluso detectó la consciencia en dos pacientes que sufrían síndrome de «enclaustramiento» y no podían moverse ni comunicarse de ninguna manera normal.[117] Así pues, esto está naciendo como una tecnología prometedora para que los médicos en el futuro puedan saber si determinados pacientes son conscientes o no.

Anclar la consciencia en la física

La TII solo está definida para sistemas discretos que pueden encontrarse en un número finito de estados, por ejemplos los bits en la memoria de un ordenador o unas neuronas sobresimplificadas que pueden estar activas o inactivas. Por desgracia, esto significa que la TII no está definida para la mayoría de los sistemas físicos tradicionales, que pueden cambiar de forma continua (por ejemplo, la posición de una partícula o la intensidad de un campo magnético pueden tomar cualquiera de entre una cantidad infinita de valores).[118] Si tratamos de aplicar la fórmula de la TII a tales sistemas, normalmente obtendremos el resultado nada útil de que Φ es infinito. Los sistemas mecano-cuánticos pueden ser discretos, pero la TII original no está definida para sistemas cuánticos. ¿Cómo podemos anclar la TII y otras teorías de la consciencia basadas en la información sobre unos sólidos cimientos físicos?

Podemos hacerlo partiendo de lo que aprendimos en el capítulo 2 sobre cómo los pedazos de materia pueden tener propiedades emergentes en relación con la información. Vimos que, para que algo pudiese usarse como un dispositivo de memoria capaz de almacenar información, debía tener muchos estados estables. También vimos que para que pudiese ser *computronio*, una sustancia susceptible de realizar computaciones, se necesitaba además una dinámica compleja: las leyes de la física debían hacer que cambiase de maneras lo suficientemente complicadas para poder implementar un procesamiento de información arbitrario. Por último, vimos cómo una red neuronal, por ejemplo, es un potente sustrato para el

aprendizaje porque, al obedecer las leyes de la física, puede reorganizarse para incrementar de forma gradual su capacidad de implementar las computaciones deseadas. Estamos planteando una pregunta adicional: ¿qué hace que un pedazo de materia pueda tener una experiencia subjetiva? O, dicho de otro modo, ¿bajo qué condiciones podrá un pedazo de materia hacer estas cuatro cosas?

1. Recordar.
2. Computar.
3. Aprender.
4. Experimentar.

Vimos las tres primeras en el capítulo 2, y ahora estamos abordando la cuarta. Igual que Margolus y Toffoli acuñaron el término *computronio* para referirse a una sustancia capaz de realizar computaciones arbitrarias, me gustaría usar el término *sentronio* para la sustancia más general que tenga una experiencia subjetiva (sea capaz de sentir).(33)

Pero ¿cómo puede la consciencia experimentarse como algo tan poco físico si se trata de hecho de un fenómeno físico? ¿Cómo puede sentirse tan independiente de su sustrato físico? Creo que se debe a que es bastante independiente de su sustrato físico, de la materia en la que está modelada. En el capítulo 2, vimos muchos ejemplos fantásticos de patrones independientes del sustrato, como las ondas, las memorias y las computaciones. Vimos cómo no solo eran más que la suma de sus partes (emergentes), sino que eran bastante independientes de dichas partes, que cobraban vida propia. Por ejemplo, vimos cómo una futura mente simulada o un personaje de videojuego no tendría forma de saber si estaba ejecutándose sobre Windows, Mac OS, un teléfono Android o algún otro sistema operativo, porque sería independiente del sustrato. Como tampoco podría saber si las puertas lógicas de su ordenador estaban hechas de transistores, circuitos ópticos u otro tipo de hardware. O cuáles son las leyes físicas fundamentales, que podrían adoptar cualquier forma siempre que permitiesen la construcción de ordenadores universales.

En resumen, creo que la consciencia es un fenómeno físico que se siente como no físico porque es como las ondas y las computaciones: tiene propiedades independientes de su sustrato físico específico. Esto es algo que se deduce lógicamente de la idea de la consciencia como información, y

conduce a un concepto radical que me gusta mucho: si la consciencia es la manera en que la información se siente cuando se procesa de determinadas maneras, entonces debe ser independiente del sustrato; lo único que importa es la estructura del procesamiento de información, no la estructura de la materia que lleva a cabo tal procesamiento. Dicho de otro modo, la consciencia es doblemente independiente del sustrato.

Como hemos visto, la física describe patrones en el espacio-tiempo que corresponden a partículas que se mueven de un sitio a otro. Si las disposiciones de partículas obedecen ciertos principios, dan lugar a fenómenos emergentes que son en buena medida independientes del sustrato de partículas, y se experimentan de una manera totalmente diferente. Un ejemplo excelente de esto último es el procesamiento de información en computronio. Pero ahora hemos llevado esta idea a otro nivel: *Si el procesamiento de información en sí obedece a ciertos principios, entonces puede dar lugar al fenómeno emergente de más alto nivel que denominamos consciencia.* Esto sitúa nuestra experiencia consciente no uno sino dos niveles por encima de la materia. ¡No es de extrañar que experimentemos nuestra mente como si no fuese física!

Esto suscita una pregunta: ¿cuáles son estos principios que el procesamiento de información debe obedecer para ser consciente? No pretendo saber qué condiciones son *suficientes* para garantizar la consciencia, pero a continuación expongo cuatro condiciones *necesarias* por las que apostaría y que he explorado en mis investigaciones:

PRINCIPIO	DEFINICIÓN
Principio de información	Un sistema consciente tiene una capacidad de almacenamiento de información considerable.
Principio de dinámica	Un sistema consciente tiene una capacidad de procesamiento de información considerable.
Principio de independencia	Un sistema consciente tiene una independencia considerable respecto del resto del mundo.
Principio de integración	Un sistema consciente no puede estar compuesto por partes casi independientes.

Como he dicho, creo que la consciencia es la manera en que se experimenta la información cuando se procesa de determinadas maneras. Esto

significa que, para ser consciente, un sistema debe ser capaz de almacenar y procesar información, lo que corresponde a los dos primeros principios. Tengamos en cuenta que la memoria no necesita durar mucho tiempo: recomiendo ver este conmovedor vídeo de Clive Wearing, que parece perfectamente consciente a pesar de que sus recuerdos duran menos de un minuto.[119] Creo que un sistema consciente también necesita ser bastante independiente del resto del mundo, porque de lo contrario no sentiría subjetivamente que tuviera existencia independiente alguna. Por último, creo que el sistema consciente debe estar integrado como un todo unificado, como argumentó Giulio Tononi, porque, si tuviera dos partes independientes, se sentirían como dos entidades conscientes distintas, en lugar de una sola. Los primeros tres principios implican *autonomía*: que el sistema es capaz de retener y procesar información sin demasiada interferencia externa, y por lo tanto determina su propio futuro. Los cuatro principios juntos significan que un sistema es autónomo, pero que sus partes no lo son.

Si estos cuatro principios son correctos, entonces está claro lo que tenemos que hacer: debemos buscar teorías matemáticamente rigurosas que los incorporen y ponerlas a prueba de forma experimental. También necesitamos determinar si se necesitan principios adicionales. Con independencia de si la TII es correcta o no, los investigadores deben tratar de desarrollar teorías alternativas y poner a prueba todas las teorías disponibles mediante experimentos cada vez mejores.

CONTROVERSIAS DE LA CONSCIENCIA

Ya hemos comentado la eterna polémica en torno a si la investigación sobre la consciencia es un disparate acientífico y una innecesaria pérdida de tiempo. Además, hay otras controversias recientes en la vanguardia de la investigación sobre la consciencia. Echemos un vistazo a las que considero más reveladoras.

En los últimos tiempos, la TII de Giulio Tononi ha recibido no solo elogios sino también críticas, algunas de ellas mordaces. Recientemente, Scott Aaronson afirmó lo siguiente en su blog: «En mi opinión, el hecho de que la teoría de la información integrada es errónea —demostrablemente errónea, por razones que afectan a su núcleo— la sitúa en algo así como el 2 %

superior de todas las teorías matemáticas de la consciencia que se han propuesto a lo largo de la historia. A mi modo de ver, casi todas las demás teorías son tan vagas, vaporosas y maleables que no pueden más que aspirar a ser erróneas».[120] En descargo tanto de Scott como de Giulio, debo decir que no llegaron a las manos cuando los vi debatir sobre la TII en un reciente seminario de la Universidad de Nueva York, y que ambos escucharon educadamente los argumentos del otro. Aaronson demostró que ciertas redes simples de puertas lógicas tenían una información integrada (Φ) muy elevada y sostuvo que, puesto que claramente no eran conscientes, la TII era errónea. Giulio replicó que, si se construyesen tales redes, *serían* conscientes, y el hecho de que Scott diese por descontado lo contrario se debía a un sesgo antropocéntrico, como si el dueño de un matadero afirmase que los animales no podían ser conscientes porque no podían hablar y eran muy distintos de los humanos. Mi análisis, con el que ambos estuvieron de acuerdo, era que tenían visiones opuestas sobre si la integración era solo una condición *necesaria* para la consciencia (algo que a Scott le parecía correcto) o también una condición *suficiente* (como afirmaba Giulio). Evidentemente, esta última es una afirmación más fuerte y polémica, que confío en que pronto podamos poner a prueba de forma experimental.[121]

Otra afirmación controvertida de la TII es que la arquitectura de los ordenadores actuales no puede ser consciente, porque la manera en que están conectadas sus puertas lógicas hace que tengan una integración muy baja.[122] En otras palabras, si nos replicamos en un futuro robot de altas prestaciones que simule con precisión todas y cada una de nuestras neuronas y sinapsis, incluso si este clon digital tiene nuestro mismo aspecto, y habla y actúa de manera que resulta indistinguible de nosotros, Giulio afirma que sería un zombi inconsciente sin experiencia subjetiva; lo cual sería decepcionante si nos hubiésemos replicado buscando la inmortalidad subjetiva.(34) Tanto David Chalmers como el profesor de IA Murray Shanahan han puesto en duda esta afirmación al imaginar lo que sucedería si en lugar de la situación anterior fuésemos sustituyendo gradualmente los circuitos neuronales en nuestro cerebro por un hipotético hardware digital que los simulase a la perfección.[123] Aunque nuestro *comportamiento* no se vería afectado por la sustitución, ya que estamos suponiendo que la simulación es perfecta, nuestra *experiencia*, según Giulio, pasaría de ser inicialmente consciente a inconsciente al final. Pero ¿qué se sentiría en los

estadios intermedios, a medida que una proporción cada vez mayor fuese sustituida? Cuando fuesen reemplazadas las partes de nuestro cerebro responsables de nuestra experiencia consciente en la mitad superior de nuestro campo visual, ¿notaríamos que parte de nuestro escenario visual de pronto ha desaparecido, pero que, a pesar de ello, misteriosamente sabíamos lo que había ahí, como cuentan los pacientes que padecen «visión ciega»? [124] Esto sería muy inquietante porque, si pudiésemos experimentar conscientemente alguna diferencia, también podríamos contársela a nuestros amigos cuando nos preguntasen. Pero estamos suponiendo que nuestro comportamiento no puede cambiar. La única posibilidad lógica compatible con las suposiciones es que, exactamente en el mismo instante en que una cosa desaparece de nuestra consciencia, nuestra mente se altera misteriosamente para hacer o bien que mintamos y neguemos que nuestra experiencia ha cambiado, o bien para que olvidemos que las cosas eran diferentes.

Por otra parte, Murray Shanahan reconoce que la misma crítica de la sustitución gradual se le puede plantear a *cualquier* teoría que afirme que podemos actuar de forma consciente sin ser conscientes, por lo que podríamos estar tentados de concluir que actuar y ser conscientes son la misma cosa, y que, por lo tanto, el comportamiento observable externamente es lo único que importa. Pero entonces habríamos caído en la trampa de predecir que no somos conscientes mientras soñamos, aunque sabemos que eso no es así.

Una tercera controversia en relación con la TII es la que gira en torno a si una entidad consciente puede estar compuesta de partes que sean conscientes por separado. Por ejemplo, ¿puede la sociedad en su conjunto cobrar consciencia sin que las personas pierdan la suya? ¿Puede un cerebro consciente tener partes que también sean conscientes por su cuenta? La predicción de la TII es un «no» rotundo, pero esto no convence a todo el mundo. Por ejemplo, algunos pacientes con lesiones que reducen severamente la comunicación entre las dos mitades de su cerebro experimentan el «síndrome de la mano ajena», por el que su cerebro derecho hace que su mano izquierda haga cosas que los pacientes afirman que ellos no están provocando o entendiendo, en ocasiones hasta el extremo de que usan su otra mano para inmovilizar su mano «ajena». ¿Cómo podemos estar tan seguros de que no hay dos consciencias distintas en su cabeza, una en el

hemisferio derecho que es incapaz de hablar y otra en el izquierdo que es la única que habla, y que afirma hacerlo en nombre de las dos? Imaginemos que usamos una tecnología futura para construir un canal de comunicación directa entre dos cerebros humanos, y que incrementamos gradualmente la capacidad de este canal hasta que la comunicación es tan eficiente entre dos cerebros como en el interior de uno de ellos. ¿Llegaría un momento en que las dos consciencias individuales desaparecerían de pronto y serían reemplazadas por una única consciencia unificada, tal y como predice la TII, o esa transición sería gradual, de manera que las consciencias individuales coexistirían en alguna forma incluso a medida que empezase a surgir una experiencia conjunta?

Otra controversia fascinante es si los experimentos subestiman de cuántas cosas somos conscientes. Vimos con anterioridad que, aunque sentimos que somos visualmente conscientes de gran cantidad de información (colores, formas, objetos y en apariencia todo lo que tenemos delante), los experimentos han demostrado que solo podemos recordar y dar cuenta de una fracción ridículamente pequeña de todo eso.[125] Algunos investigadores han intentado resolver esta discrepancia preguntándonos si a veces podemos tener «consciencia sin acceso», es decir, experiencia subjetiva de cosas que son demasiado complejas para caber en nuestra memoria de trabajo para su uso posterior.[126] Por ejemplo, cuando experimentamos *ceguera por falta de atención* por estar demasiado distraídos para percatarnos de un objeto a simple vista, esto no implica que no tengamos una experiencia visual consciente de ello, simplemente que no se ha almacenado en nuestra memoria de trabajo.[127] ¿Debemos interpretarlo como olvido más que como ceguera? Otros investigadores rechazan esta idea de que no se puede confiar en lo que las personas dicen que experimentaron y advierten sobre sus implicaciones. Murray Shanahan imagina un ensayo clínico donde los pacientes dan fe del alivio completo del dolor gracias a un nuevo fármaco maravilloso que, sin embargo, es rechazado por una comisión gubernamental: «Los pacientes solo creen que no sienten dolor. Gracias a la neurociencia, sabemos que no es así».[128] Por otro lado, ha habido casos en los que a los pacientes que despertaron accidentalmente durante la cirugía se les administró un fármaco para que olvidaran el mal trago. ¿Debemos confiar en su testimonio posterior de que no experimentaron dolor?[129]

¿CÓMO SE SENTIRÍA LA CONSCIENCIA DE UNA ia?

Si algún futuro sistema de IA es consciente, ¿qué experimentará subjetivamente? Esta es la esencia del «problema aún más difícil» de la consciencia, y nos obliga a ascender hasta el segundo nivel de dificultad que se muestra en la figura 8.1. No solo carecemos hoy en día de una teoría que responda a esta pregunta, sino que ni siquiera estamos seguros de si es lógicamente posible dar una respuesta completa. Al fin y al cabo, ¿cuál podría ser una respuesta satisfactoria? ¿De qué manera le explicaríamos a una persona ciega cómo es el color rojo?

Por suerte, nuestra incapacidad actual de ofrecer una respuesta completa no nos impide dar respuestas parciales. Unos extraterrestres inteligentes que estudiasen el sistema sensorial humano probablemente deducirían que los colores son los qualia que se sienten asociados con cada punto de una superficie bidimensional (nuestro campo visual), mientras que los sonidos no se sienten localizados espacialmente, y los dolores son qualia que se sienten asociados con distintas partes del cuerpo. A partir del descubrimiento de que nuestras retinas poseen tres tipos de células cónicas sensibles a la luz, podrían inferir que experimentamos tres colores primarios y que todos los demás qualia de color resultan de la combinación de dichos colores primarios. Si midieran cuánto tiempo tardan las neuronas en transmitir información a través del cerebro, podrían concluir que no experimentamos más de diez pensamientos o percepciones conscientes por segundo, y que, cuando vemos películas en la televisión a veinticuatro fotogramas por segundo, no experimentamos una sucesión de imágenes estáticas sino un movimiento continuo. Cuando consiguieran medir la velocidad a la que se libera adrenalina en nuestro flujo sanguíneo y cuánto tiempo permanece en él antes de descomponerse, podrían predecir que sentimos arranques de ira que comienzan al cabo de unos pocos segundos y pueden durar varios minutos.

Aplicando argumentos similares basados en la física, podemos hacer conjeturas más o menos fundamentadas sobre determinados aspectos de cómo podría sentirse una consciencia artificial. En primer lugar, el espacio de todas las posibles experiencias de IA es *inmenso* comparado con lo que los humanos podemos experimentar. Nosotros tenemos una clase de qualia por

cada uno de nuestros sentidos, pero las IA pueden tener una cantidad muchísimo mayor de tipos de sensores y de representaciones internas de la información, por lo que hemos de evitar caer en el error de suponer que ser una IA necesariamente se siente como algo parecido a ser una persona.

En segundo lugar, una consciencia artificial del tamaño de un cerebro podría tener millones de veces más experiencias que nosotros por segundo, puesto que las señales electromagnéticas viajan a la velocidad de la luz, millones de veces más rápido que las señales neuronales. Sin embargo, cuanto mayor fuese la IA, más lentos tendrán que ser sus pensamientos para que la información tenga tiempo de fluir entre todas sus partes, como vimos en el capítulo 4. Así, cabría esperar que una IA «Gaia» del tamaño de la Tierra solo tuviese unas diez experiencias conscientes por segundo, como un humano, y que una IA del tamaño de una galaxia solo pudiese tener un pensamiento global aproximadamente cada 100.000 años (por lo tanto, no más de unas cien experiencias durante toda la historia del universo transcurrida hasta ahora). Esto daría a las IA grandes motivos aparentemente irresistibles para delegar computaciones en los sistemas más pequeños capaces de realizarlas, para acelerar las cosas, igual que nuestra mente consciente ha delegado el parpadeo reflejo a un subsistema pequeño, rápido e inconsciente. Aunque vimos antes que el procesamiento consciente de información en nuestros cerebros parece ser solamente la punta de un iceberg, por lo demás inconsciente, debemos esperar que la situación sea aún más extrema para las IA grandes futuras: si tienen una única consciencia, entonces es probable que no sea consciente de casi todo el procesamiento de información que tiene lugar en su interior. Es más, aunque las experiencias conscientes que disfruta pueden ser sumamente complejas, también serán lentísimas en comparación con las veloces actividades de sus partes más pequeñas.

Esto pone realmente de relieve la controversia antes mencionada sobre si partes de una entidad consciente también pueden ser conscientes. La TII predice que no, lo que significa que si una IA de tamaño astronómico es consciente, entonces casi todo su procesamiento de información será inconsciente. Esto significaría que, si una civilización de IA más pequeñas mejora sus capacidades de comunicación hasta el punto en que surge una sola mente colectiva consciente, de pronto sus consciencias individuales, mucho más rápidas, se apagarían. Por otro lado, si la predicción de la TII es errónea,

la mente colectiva puede coexistir con multitud de mentes conscientes más pequeñas. De hecho, uno podría incluso imaginar una jerarquía anidada de consciencias a todos los niveles, desde las microscópicas hasta las cósmicas.

Como vimos anteriormente, el procesamiento de información inconsciente en el cerebro humano parece estar relacionado con la manera de pensar sin esfuerzo, rápida y automática que los psicólogos llaman «sistema 1».[130] Por ejemplo, nuestro sistema 1 podría informar a la consciencia de que su análisis sumamente complejo de la información visual ha determinado que ha llegado nuestro mejor amigo, sin darnos ninguna pista de cómo se realizó la computación. Si esta relación entre los sistemas y la consciencia resulta ser válida, será tentador generalizar esta terminología a las IA, y referirse a todas las tareas rutinarias rápidas delegadas a subunidades inconscientes como el sistema 1 de la IA. El pensamiento global lento, costoso y controlado de la IA sería, si fuera consciente, el sistema 2 de la IA. Los humanos también tenemos experiencias conscientes en las que interviene lo que llamaré el «sistema 0»: percepción pasiva sin procesar que tiene lugar incluso cuando permanecemos sin movernos y sin pensar y nos limitamos a observar el mundo que nos rodea. Los sistemas 0, 1 y 2 parecen progresivamente más complejos, por lo que es sorprendente que solo el del medio parezca ser inconsciente. La TII explica esto diciendo que la información sensorial sin procesar del sistema 0 se almacena en estructuras cerebrales parecidas a una cuadrícula con una integración muy alta, mientras que el sistema 2 tiene una alta integración debido a circuitos de retroalimentación, mediante los que toda la información de la que somos conscientes ahora puede afectar nuestros futuros estados cerebrales. Por otra parte, fue precisamente la predicción de la red consciente la que desencadenó la crítica de Scott Aaronson a la TII. En resumen, si una teoría que resuelve el problema bastante difícil de la consciencia logra algún día superar una batería rigurosa de pruebas experimentales para que comencemos a tomar en serio sus predicciones, entonces también reducirá enormemente las opciones para el problema aún más difícil en cuanto a lo que las IA futuras conscientes podrían experimentar.

Algunos aspectos de nuestra experiencia subjetiva se remontan claramente a nuestros orígenes evolutivos, por ejemplo nuestros deseos emocionales relacionados con la conservación (comer, beber, evitar que nos maten) y la reproducción. Esto significa que debería ser posible crear IA que nunca

experimente qualia como el hambre, la sed, el miedo o el deseo sexual. Como vimos en el capítulo anterior, si se programa una IA muy inteligente para que tenga prácticamente cualquier objetivo ambicioso, es probable que se esfuerce por conservarse para poder alcanzar dicho objetivo. Sin embargo, si forman parte de una sociedad de IA, puede que carezcan de nuestro intenso miedo a la muerte: si disponen de réplicas de sí mismas, todo lo que se arriesgan a perder son los recuerdos que hayan acumulado desde que se produjo la copia de respaldo más reciente, siempre que confíen en que esta se usará. Además, la capacidad de copiar con facilidad información y software entre IA probablemente rebajaría el fuerte sentimiento de individualidad tan característico de nuestra consciencia humana: habría una menor distinción entre usted y yo si pudiésemos compartir y copiar todos nuestros recuerdos y capacidades, por lo que un grupo de IA cercanas podrían sentirse más como un solo organismo con una mente colectiva.

¿Sentiría una consciencia artificial que tiene libre albedrío? Cabe señalar que, aunque los filósofos llevan miles de años discutiendo sobre si *nosotros* tenemos libre albedrío sin alcanzar ningún consenso ni siquiera sobre cómo definir la cuestión,[131] la pregunta que yo planteo es diferente, y posiblemente más fácil de abordar. Permítame que intente convencerlo de que la respuesta es: «Sí, cualquier entidad consciente que deba tomar decisiones, tanto si es biológica como artificial, *sentirá* subjetivamente que tiene libre albedrío». Las decisiones se sitúan en un espectro entre dos extremos:

1. Sabemos exactamente por qué tomamos esa decisión en particular.
2. No tenemos ni idea de por qué tomamos esa decisión en particular; sentimos como si lo hubiésemos hecho aleatoriamente por capricho.

Las discusiones sobre el libre albedrío suelen girar en torno a las dificultades para reconciliar nuestro comportamiento intencional de toma de decisiones con las leyes de la física: si elegimos entre las dos siguientes explicaciones para lo que hemos hecho, ¿cuál de ellas es la correcta: «Le pedí salir porque me gustaba mucho» o «Mis partículas hicieron que le pidiese salir moviéndose de acuerdo con las leyes de la física»? Pero vimos en el capítulo anterior que *ambas* lo son: lo que se siente como comportamiento intencional puede surgir de leyes físicas deterministas y ciegas. Más concretamente,

cuando un sistema (cerebro o IA) toma una decisión de tipo 1, la computa usando algún algoritmo determinista, y la razón por la que siente que ha tomado una decisión es que de hecho lo hizo cuando computó qué hacer. Es más, como señala Seth Lloyd,[132] hay un famoso teorema informático que afirma que, para casi todas las computaciones, no existe modo más rápido de determinar cuál será su resultado que ejecutarlas de manera efectiva. Esto significa que normalmente es imposible determinar lo que decidiremos dentro de un segundo en menos de un segundo, lo que contribuye a reforzar nuestra experiencia de tener libre albedrío. Por el contrario, cuando un sistema (cerebro o IA) toma una decisión de tipo 2, simplemente programa su mente para que base su decisión en el resultado de algún subsistema que actúa como un generador de números aleatorios. Tanto en un cerebro como en un ordenador, los números aleatorios se generan fácilmente amplificando el ruido. Por lo tanto, con independencia de dónde se sitúe una decisión en el espectro entre la opción 1 y la 2, tanto las consciencias biológicas como las artificiales pueden sentir que tienen libre albedrío: sienten que son realmente ellas las que deciden y no pueden predecir con certeza cuál será la decisión hasta que han terminado de pensarla.

Hay gente que me dice que la causalidad les parece degradante, que hace que sus procesos de pensamiento carezcan de sentido y que los convierte en meras «máquinas». Tal negatividad me parece absurda e injustificada. Para empezar, los cerebros humanos no tienen nada de «meros»; por lo que a mí respecta, son los objetos más asombrosamente sofisticados en el universo conocido. En segundo lugar, ¿qué alternativa preferiría esta gente? ¿Acaso no quieren que sean sus propios procesos mentales (las computaciones efectuadas por sus cerebros) los que tomen sus decisiones? Su experiencia subjetiva de libre albedrío es simplemente cómo sus computaciones se sienten desde dentro: no saben el resultado de una computación hasta que la han completado. Esto es lo que significa decir que la computación es la decisión.

SENTIDO

Para finalizar, volvamos al punto de partida de este libro: ¿cómo queremos que sea el futuro de la vida? Vimos en el capítulo anterior cómo distintas

culturas de todo el mundo aspiran a un futuro repleto de experiencias positivas, pero que surgen controversias fascinantemente espinosas cuando se intenta encontrar un consenso sobre lo que debe considerarse positivo y cómo alcanzar equilibrios entre lo que es bueno para las distintas formas de vida. Pero no dejemos que esas polémicas nos distraigan de lo esencial: no puede haber experiencias positivas si no hay experiencias, esto es, si no hay consciencia. Dicho de otra manera, sin consciencia no puede haber felicidad, bondad, belleza, sentido o propósito; solamente un astronómico desperdicio de espacio. Esto implica que, cuando la gente pregunta por el significado de la vida como si el cosmos tuviese la obligación de darle sentido a nuestra existencia, están entendiendo la situación al revés: *No es el universo el que da sentido a los seres conscientes, sino los seres conscientes los que dan sentido al universo*. De manera que nuestro primerísimo objetivo en la lista de deseos para el futuro debe ser conservar (y, con suerte, ampliar) la consciencia biológica y/o artificial en el cosmos, en lugar de abocarla a la extinción.

Si tenemos éxito en esta tarea, ¿cómo nos sentiremos los humanos al convivir con máquinas cada vez más inteligentes? ¿Le molesta a usted la irrupción aparentemente inexorable de la inteligencia artificial? Si es así, ¿por qué? En el capítulo 3 vimos cómo debería ser relativamente fácil para una tecnología basada en IA satisfacer nuestras necesidades básicas, como las de seguridad e ingresos, siempre que exista la voluntad política de hacerlo. Sin embargo, quizá le inquiete pensar que estar bien alimentado, vestido, alojado y entretenido no sea suficiente. Si nos garantizan que la IA se encargará de cubrir todas nuestras necesidades prácticas y todos nuestros deseos, ¿podríamos, no obstante, acabar sintiendo una carencia de sentido y propósito en nuestras vidas, como si fuéramos animales de zoológico bien cuidados?

Tradicionalmente, los humanos hemos basado muchas veces nuestra autoestima en la idea del *excepcionalismo humano*: la convicción de que somos las entidades más inteligentes del planeta, y por lo tanto especiales y superiores. La irrupción de la IA nos obligará a abandonar estas ideas y a ser más humildes. Pero quizá esto sea algo que debemos hacer de todas formas: a fin de cuentas, aferrarse a arrogantes ideas de superioridad sobre otros (individuos, grupos étnicos, especies, etcétera) ha provocado espantosos problemas en el pasado. De hecho, el excepcionalismo humano no solo ha provocado dolor en el pasado, sino que también parece innecesario para la

prosperidad de la humanidad: si descubrimos una civilización extraterrestre pacífica y mucho más avanzada que nosotros en ciencia, arte y todo lo demás que nos importa, presumiblemente eso no impediría que las personas siguiesen encontrándoles sentido y propósito a sus vidas. Con suerte, podríamos seguir teniendo nuestras familias, amigos y comunidades, y todas las actividades que nos proporcionan sentido y propósito, y no perder más que la arrogancia.

Mientras planificamos nuestro futuro, consideremos el significado no solo de nuestras propias vidas, sino también del propio universo. Aquí, dos de mis físicos favoritos, Steven Weinberg y Freeman Dyson, representan puntos de vista diametralmente opuestos. Weinberg, que ganó el Premio Nobel por su trabajo seminal sobre el modelo estándar de la física de partículas, dijo: «Cuanto más comprensible nos parece el universo, menos sentido parece tener».[133] Dyson, por su parte, es mucho más optimista, como ya vimos en el capítulo 6: aunque está de acuerdo en que el universo era absurdo, cree que ahora la vida le está confiriendo cada vez más sentido, y que lo mejor está aún por llegar si la vida logra difundirse por todo el cosmos. Dyson terminó su artículo seminal de 1979 diciendo: «¿Está el universo de Weinberg o el mío más cerca de la verdad? Algún día, dentro de poco tiempo, lo sabremos».[134] Si el universo vuelve a un estado permanente de inconsciencia porque provocamos la extinción de la vida en la Tierra o porque permitimos que una IA zombi e inconsciente se haga con el control del cosmos, eso dará a Weinberg toda la razón.

Desde esta perspectiva, vemos que, aunque en este libro nos hemos centrado en el futuro de la inteligencia, el futuro de la consciencia es aún más importante, ya que esta última es la que hace posible el sentido. A los filósofos les gusta hablar en latín sobre esta distinción, y contraponer la *sapiencia* (la capacidad de pensar de forma inteligente) con la *sentiencia* (la capacidad de experimentar subjetivamente qualia). Los humanos hemos construido nuestra identidad en torno a ser *Homo sapiens*, las entidades más inteligentes que existen. Mientras nos preparamos para recibir una lección de humildad de máquinas cada vez más inteligentes, sugiero que pasemos a identificarnos como *Homo sentiens*.

- No hay una definición indiscutible de «consciencia». Utilizo la definición amplia y no antropocéntrica de «consciencia = *experiencia subjetiva*».
- Saber si las IA son conscientes o no en ese sentido es lo que importa de cara a los problemas éticos y filosóficos más espinosos que plantea la irrupción de la IA: ¿pueden las IA sufrir? ¿Deben tener derechos? ¿Transmigrar un alma a una réplica digital es un suicidio subjetivo? ¿Podría un futuro cosmos repleto de IA ser un gigantesco apocalipsis zombi?
- El problema de entender la inteligencia no debe confundirse con otros tres relacionados con la consciencia: el «problema bastante difícil» de predecir qué sistemas físicos son conscientes, el «problema aún más difícil» de predecir los qualia y el «problema realmente difícil» de por qué algo es consciente.
- El «problema bastante difícil» de la consciencia es científico, ya que una teoría que predice qué procesos cerebrales son conscientes es experimentalmente comprobable y falsable, mientras que hoy en día no está claro cómo podría la ciencia resolver del todo los otros dos problemas más difíciles.
- Los avances en neurociencia sugieren que muchos comportamientos y regiones cerebrales son inconscientes, y gran parte de nuestra experiencia consciente es un resumen *a posteriori* de cantidades mucho más grandes de información inconsciente.
- Para poder generalizar las predicciones sobre la consciencia de los cerebros a las máquinas se necesita una teoría. La consciencia no parece requerir un cierto tipo de partícula o campo, sino un modo de procesamiento de información que sea relativamente autónomo e integrado, de modo que el sistema en su conjunto sea relativamente autónomo, pero sus partes no.
- La consciencia puede sentirse como algo muy poco físico porque es doblemente independiente del sustrato: si la consciencia es la forma en que se siente la información cuando se procesa de determinadas maneras complejas, entonces lo que importa es simplemente la estructura del procesamiento de la información, no la estructura de la materia que la procesa.
- Si es posible la consciencia artificial, entonces es probable que el espacio de posibles experiencias de IA sea enorme en comparación con lo que los humanos experimentamos, que abarquen un amplio espectro de qualia y de escalas temporales y que compartan todas ellas la sensación de tener libre albedrío.
- Como no puede haber significado sin consciencia, no es el universo el que da sentido a los seres conscientes, sino los seres conscientes los que dan sentido al universo.
- Esto sugiere que, ahora que nos preparamos para recibir una lección de humildad de máquinas cada vez más inteligentes, sería conveniente que nos acostumbrásemos a identificarnos sobre todo como *Homo sentiens*, no como *Homo sapiens*.

EPÍLOGO
LA HISTORIA DEL EQUIPO DEL FLI

El aspecto más triste de la vida actual es que la ciencia gana en conocimiento más rápidamente que la sociedad en sabiduría.

ISAAC ASIMOV

Henos aquí, querido lector, al final del libro, tras haber explorado el origen y el destino de la inteligencia, los objetivos y el sentido. ¿Cómo podemos llevar estas ideas a la práctica? ¿Qué debemos hacer en concreto para que nuestro futuro sea lo mejor posible? Esta es precisamente la pregunta que me planteo en este momento, sentado junto a la ventana del avión que me lleva de vuelta a Boston desde San Francisco el 9 de enero de 2017, tras haber participado en la conferencia que acabamos de organizar en Asilomar. Como cierre del libro, permítame compartir mis pensamientos con usted.

Meia está a mi lado, recuperando el sueño perdido tras muchas largas noches preparando y organizando la conferencia. ¡Qué locura de semana! Conseguimos reunir a casi todas las personas que he mencionado en este libro durante unos pocos días en esta continuación de la conferencia de Puerto Rico, entre ellas emprendedores como Elon Musk y Larry Page, destacados investigadores en IA, tanto del mundo académico como de empresas como DeepMind, Google, Facebook, Apple, IBM, Microsoft y Baidu, así como economistas, expertos en derecho, filósofos y otros extraordinarios pensadores (véase la figura 9.1). Los resultados superaron incluso mis ambiciosas expectativas, y me siento más optimista sobre el futuro de la vida de lo que lo he estado en mucho tiempo. En este epílogo, le voy a explicar por qué.

HA NACIDO EL FLI

Desde que, a los catorce años, supe de la carrera armamentística nuclear, siempre me ha preocupado el hecho de que el poder de nuestra tecnología aumentaba más rápidamente que la sabiduría con la que la gestionamos. Por eso, decidí introducir un capítulo sobre este problema en mi primer libro, *Nuestro universo matemático*, aunque el resto del mismo trataba principalmente sobre física. Para 2014, uno de mis propósitos de Año Nuevo fue el de no permitirme a mí mismo quejarme sobre algo sin antes pensar seriamente qué podía hacer yo al respecto, y mantuve mi promesa durante mi gira de presentación del libro ese mes de enero: Meia y yo le dimos muchas vueltas a la idea de cómo lanzar una especie de organización sin ánimo de lucro centrada en mejorar el futuro de la vida a través de la buena gestión de la tecnología.



FIGURA 9.1. Nuestra conferencia de enero de 2017 en Asilomar, continuación de la de Puerto Rico, reunió a un extraordinario grupo de investigadores en IA y otros campos relacionados. Segunda fila, de izquierda a derecha: Patrick Lin, Daniel Weld, Ariel Conn, Nancy Chang, Tom Mitchell, Ray Kurzweil, Daniel Dewey, Margaret Boden, Peter Norvig, Nick Hay, Moshe Vardi, Scott Siskind, Nick Bostrom, Francesca Rossi, Shane Legg, Manuela Veloso, David Marble, Katja Grace, Irakli Beridze, Marty Tenenbaum, Gill Pratt, Martin Rees, Joshua Greene, Matt Scherer, Angela Kane, Amara Angelica, Jeff Mohr, Mustafa Suleyman, Steve Omohundro, Kate Crawford, Vitalik Buterin, Yutaka Matsuo, Stefano Ermon, Michael Wellman, Bas Steunebrink, Wendell Wallach, Allan Dafoe, Toby Ord, Thomas Dietterich, Daniel Kahneman, Dario Amodei, Eric Drexler, Tomaso Poggio, Eric Schmidt, Pedro Ortega, David Leake, Seán Ó hÉigeartaigh, Owain Evans, Jaan Tallinn, Anca Dragan, Sean Legassick, Toby Walsh, Peter Asaro, Kay Firth-Butterfield, Philip Sabes, Paul Merolla, Bart Selman, Tucker Davey, ?, Jacob Steinhardt, Moshe Looks, Josh Tenenbaum, Tom Gruber, Andrew Ng, Kareem Ayoub, Craig Calhoun, Percy Liang, Helen Toner, David Chalmers, Richard Sutton, Claudia Passos-Ferreira, János Krámar, William MacAskill, Eliezer Yudkowsky, Brian Ziebart, Huw Price, Carl Shulman, Neil Lawrence, Richard Mallah, Jurgen Schmidhuber, Dileep George, Jonathan Rothberg, Noah Rothberg. Primera fila: Anthony Aguirre, Sonia Sachs, Lucas Perry, Jeffrey Sachs, Vincent Conitzer, Steve Goose, Victoria Krakovna, Owen Cotton-Barratt, Daniela Rus, Dylan

Hadfield-Menell, Verity Harding, Shivon Zilis, Laurent Orseau, Ramana Kumar, Nate Soares, Andrew McAfee, Jack Clark, Anna Salamon, Long Ouyang, Andrew Critch, Paul Christiano, Yoshua Bengio, David Sanford, Catherine Olsson, Jessica Taylor, Martina Kunz, Kristinn Thorisson, Stuart Armstrong, Yann LeCun, Alexander Tamas, Roman Yampolskiy, Marin Soljačić, Lawrence Krauss, Stuart Russell, Eric Brynjolfsson, Ryan Calo, ShaoLan Hsueh, Meia Chita-Tegmark, Kent Walker, Heather Roff, Meredith Whittaker, Max Tegmark, Adrian Weller, José Hernández-Orallo, Andrew Maynard, John Hering, Abram Demski, Nicolas Berggruen, Gregory Bonnet, Sam Harris, Tim Hwang, Andrew Snyder-Beattie, Marta Halina, Sebastian Farquhar, Stephen Cave, Jan Leike, Tasha McCauley, Joseph Gordon-Levitt. Llegaron después: Guru Banavar, Demis Hassabis, Rao Kambhampati, Elon Musk, Larry Page, Anthony Romero.

Meia insistió en que le pusiésemos un nombre positivo, lo más lejos posible de «Instituto de la Miseria y la Desolación» o «Instituto de la Preocupación sobre el Futuro». Puesto que el Future of Humanity Institute ya estaba cogido, acabamos llegando a Future of Life Institute (FLI), que tenía la ventaja añadida de ser más incluyente. El 22 de enero, la gira del libro nos llevó a Santa Cruz y, mientras el sol de California se ponía sobre el Pacífico, disfrutamos de la cena en compañía de nuestro viejo amigo Anthony Aguirre, al tiempo que tratábamos de convencerlo para que se uniese a nuestro proyecto. Anthony no es solo una de las personas más sabias e idealistas que conozco, sino también alguien que ha sido capaz de soportar el tener que gestionar conmigo durante más de una década otra organización sin ánimo de lucro, el Foundational Questions Institute (véase <http://fqxi.org>).

La semana siguiente, la gira hizo parada en Londres. Dado que el futuro de la IA estaba muy presente en mi mente, contacté con Demis Hassabis, quien tuvo la amabilidad de invitarme a visitar la sede de DeepMind. Me impresionó lo mucho que habían crecido desde que Demis me visitó en el MIT dos años antes. Google acababa de comprarlos por unos 650 millones de dólares, y, al ver su inmenso espacio de oficinas repleto de mentes brillantes que perseguían el audaz objetivo que Demis se había marcado de «resolver la inteligencia», tuve la sensación visceral de que cabía realmente la posibilidad de que lo lograsen.

La noche siguiente, hablé con mi amigo Jaan Tallinn a través de Skype, el software que él había ayudado a crear. Le expliqué nuestra visión para el FLI y, una hora más tarde, Jaan había decidido apostar por nosotros, y ¡financiarnos con hasta 100.000 dólares al año! Pocas cosas me emocionan más que cuando alguien deposita en mí más confianza de la que me he ganado, así que significó mucho para mí que un año más tarde, después de la

conferencia en Puerto Rico que mencioné en el capítulo 1, Jaan bromease diciendo que esta era la mejor inversión que había hecho nunca.

Al día siguiente, mi editorial había dejado un hueco en mi agenda, que aproveché para visitar el Museo de Ciencias de Londres. Tras tanto tiempo obsesionado con el pasado y el futuro de la inteligencia artificial, de pronto sentí que estaba recorriendo una manifestación física de mis pensamientos. Habían reunido una fantástica colección de objetos que representaban cómo hemos ido acumulando conocimiento, desde la locomotora Rocket de Stephenson hasta el Ford Modelo T, una réplica de tamaño real del módulo lunar Apollo 11 y ordenadores de distintas épocas, desde la calculadora mecánica que era la «máquina diferencial» de Babbage hasta aparatos actuales. También tenían una exposición sobre la historia de nuestra comprensión de la mente, desde el experimento de Galvano con ancas de rana hasta las neuronas, la electroencefalografía (EEG) y la imagen por resonancia magnética funcional (IRMf).

No lloro casi nunca, pero eso es lo que hice mientras salía de allí (en un túnel lleno de viandantes, nada menos, de camino hacia la estación de metro de South Kensington). Ahí estaban todas esas personas viviendo tranquilamente sus vidas, ajenas por completo a lo que yo estaba pensando. Primero, los humanos descubrimos cómo replicar algunos procesos naturales con máquinas, y aprendimos a producir nuestros viento y rayos, y nuestra propia potencia mecánica. De forma gradual, comenzamos a darnos cuenta de que nuestros cuerpos también eran máquinas. Posteriormente, el descubrimiento de las células nerviosas empezó a difuminar los límites entre el cuerpo y la mente. Entonces comenzamos a construir máquinas que pudiesen superar no solo a nuestros músculos, sino también a nuestras mentes. Así pues, a medida que vamos descubriendo lo que somos, ¿inevitablemente nos volvemos obsoletos a nosotros mismos? Eso sería poéticamente trágico.

Esta idea me aterraba, pero también reforzó mi determinación de cumplir mi propósito de Año Nuevo. Tenía la sensación de que nos faltaba una persona para completar el equipo de fundadores del FLI, que liderarían a un equipo de voluntarios jóvenes e idealistas. La elección lógica era Viktoriya Krakovna, una brillante estudiante de doctorado en Harvard que no solo había logrado la medalla de plata en la Olimpiada Internacional de Matemáticas, sino que además había fundado la Citadel, un hogar para

alrededor de una docena de jóvenes idealistas que querían que la razón tuviese un papel más destacado en sus vidas y en el mundo. Meia y yo invitamos a Viktoriya a nuestra casa cinco días después para exponerle nuestra idea y, antes de que hubiésemos dado cuenta del sushi, ya había nacido el FLI.

LA AVENTURA DE PUERTO RICO

Esto marcó el comienzo de una fantástica aventura, que aún continúa. Como mencioné en el capítulo 1, organizamos regularmente sesiones de lluvia de ideas en nuestra casa con decenas de estudiantes idealistas, profesores y otros pensadores locales, en las que las ideas mejor valoradas se convertían en proyectos (el primero de los cuales fue ese artículo sobre IA con Stephen Hawking, Stuart Russell y Frank Wilczek que mencioné en el capítulo 1, que contribuyó a fomentar el debate público). En paralelo con los pequeños pasos necesarios para crear una nueva organización (constituirla, reclutar a los miembros del consejo asesor, contratar al personal y lanzar el sitio web), organizamos un divertido evento para recaudar fondos ante un abarrotado auditorio del MIT, en el que Alan Alda moderó una conversación sobre el futuro de la tecnología con varios expertos destacados.



FIGURA 9.2. Jaan Tallinn, Anthony Aguirre, un servidor, Meia Chita-Tegmark y Viktoriya Krakovna celebramos con sushi la constitución del FLI el 23 de mayo de 2014.

Durante el resto de ese año, concentramos nuestros esfuerzos en sacar adelante la conferencia de Puerto Rico que, como mencioné en el capítulo 1, trataba de animar a los más destacados investigadores en IA de todo el mundo a que participasen en la conversación sobre cómo hacer que la IA fuese beneficiosa. Nuestro objetivo era llevar el debate sobre la seguridad en IA de la mera preocupación al trabajo práctico: de las disputas sobre cuán preocupados debíamos estar, a ponernos de acuerdo en torno a proyectos de investigación concretos que podrían iniciarse de inmediato para maximizar las posibilidades de que llegaran a buen puerto. Para prepararnos, recopilamos ideas prometedoras sobre investigación en torno a la seguridad en IA procedentes de todo el mundo y pedimos a la comunidad que nos diese su opinión sobre nuestra creciente lista de proyectos. Con la ayuda de Stuart Russell y de un grupo de voluntarios jóvenes y muy trabajadores, especialmente Daniel Dewey, János Krámar y Richard Mallah, resumimos las prioridades de investigación en un documento que se discutiría en la conferencia.[135] Confiábamos en que generar consenso en torno a la idea de que había muchas investigaciones valiosas que hacer sobre la IA segura propiciaría que las personas se animasen a comenzar a realizar tales

investigaciones. Nuestro éxito sería total si pudiéramos convencer a alguien de que las financiara, ya que, hasta entonces, no habían recibido apoyo de fondos gubernamentales.

Fue entonces cuando entró en escena Elon Musk. El 2 de agosto, apareció en nuestro radar con un famoso tuit: «Merece la pena leer Superinteligencia de Bostrom. Debemos ser muy prudentes con la IA. Potencialmente más peligrosa que las bombas nucleares». Traté de contactar con Elon para explicarle nuestros proyectos y pude hablar con él por teléfono unas semanas más tarde. Aunque estaba bastante nervioso y deslumbrado, el resultado fue sobresaliente: aceptó unirse al consejo científico asesor del FLI, asistir a nuestra conferencia y posiblemente financiar un primer programa de investigación en IA segura que se anunciaría en Puerto Rico. Esto nos galvanizó a todos en el FLI, e hizo que redoblásemos nuestros esfuerzos para organizar una conferencia fantástica, identificar temas de investigación prometedores y generar apoyo para ellos entre la comunidad.

Por fin pude conocer a Elon en persona para continuar con la planificación cuando vino al MIT dos meses más tarde con ocasión de un simposio sobre el espacio. Se me hizo muy raro estar a solas con él en un pequeño camerino momentos después de que hubiese embelesado a más de mil estudiantes del MIT, como si se tratase de una estrella de rock, pero al cabo de unos minutos lo único en lo que podía pensar era en nuestro proyecto conjunto. Me cayó bien desde el primer momento. Irradiaba sinceridad, y me impresionó lo mucho que le importaba el futuro a largo plazo de la humanidad, y la audacia con la que había sabido llevar sus deseos a la práctica. Quería que la humanidad explorara y colonizara el universo, y por eso creó una empresa aeroespacial. Quería energía sostenible, de manera que fundó una compañía de energía solar y otra de automóviles eléctricos. Alto, guapo, elocuente y extraordinariamente bien informado, era fácil entender por qué la gente le prestaba atención.

Desafortunadamente, este acto en el MIT también me enseñó cómo los medios de comunicación pueden dejarse llevar por el miedo y las divisiones que pueden generar. La presentación de Elon consistió en una hora de fascinante discusión sobre la exploración espacial, que creo que habría sido un estupendo programa de televisión. Al final, un alumno le hizo una pregunta sobre IA, sin relación alguna con lo que Elon había contado. Su respuesta incluyó la frase «con la inteligencia artificial, estamos convocando

al demonio», que fue lo único de lo que informaron la mayoría de los medios (y, por lo general, sacando la frase de contexto). Me llamó la atención que muchos periodistas estaban haciendo sin darse cuenta *exactamente lo contrario* de lo que queríamos conseguir en Puerto Rico. Mientras que lo que nosotros queríamos era construir consenso en la comunidad, al poner de relieve aquello en lo que estábamos de acuerdo, los medios sintieron la necesidad de destacar las divisiones. Cuanta más polémica generen, mayores serán sus registros de audiencia y sus ingresos publicitarios. Además, aunque queríamos ayudar a que personas con toda clase de opiniones encontrasen sus puntos en común, se llevasen bien y se entendiesen mejor las unas a las otras, la cobertura de los medios consiguió que personas de todo el espectro de opiniones se enfadaran entre sí, al alimentar los malentendidos publicando solo sus frases más provocativas fuera de contexto. Por esta razón, decidimos vetar la asistencia de los periodistas a la conferencia de Puerto Rico, e impusimos la «regla de Chatham House», que prohíbe a los participantes revelar posteriormente quién dijo qué.(35)

Aunque nuestra conferencia de Puerto Rico acabó siendo un éxito, conseguir que lo fuera no resultó fácil. El trabajo previo consistió sobre todo en realizar diligentemente tareas administrativas, como por ejemplo, llamar por teléfono o por Skype a una gran cantidad de investigadores en IA para alcanzar una masa crítica de participantes con la que atraer a los demás asistentes. Pero también hubo momentos dramáticos, como cuando me levanté a las siete de la mañana el 27 de diciembre para hablar con Elon, que estaba en Uruguay, a través de una pésima conexión telefónica, y oír que este me decía: «Creo que la cosa no va a funcionar». Temía que un programa de investigación en IA segura proporcionase una falsa sensación de seguridad, y eso permitiese que investigadores irresponsables siguiesen adelante comprometiéndose con la seguridad solo de palabra. Pero entonces, a pesar de que el sonido se entrecortaba continuamente, hablamos largo y tendido sobre los enormes beneficios de llamar la atención del gran público sobre el asunto y de conseguir que más investigadores trabajen en IA segura. Cuando finalizó la conversación, me envió uno de los mensajes de correo que más me ha gustado recibir: «Se cortó la llamada al final. En cualquier caso, el documento tiene buena pinta. Estaré encantado de apoyar la investigación con cinco millones de dólares en tres años. ¿Quizá deberían ser diez millones?».

Cuatro días más tarde, 2015 empezó con buen pie para Meia y para mí, que pudimos disfrutar de un breve momento de relax antes de la conferencia, bailando en Nochevieja en una playa puertorriqueña iluminada por fuegos artificiales. La conferencia también tuvo un estupendo arranque: hubo un notable acuerdo en que se necesitaba más investigación en IA segura, y, en base a los comentarios que hicieron los participantes en la conferencia, se mejoró y se completó el documento sobre líneas de investigación prioritarias. Repartimos la carta abierta en apoyo de la investigación en IA segura del capítulo 1, y quedamos encantados de que casi todo el mundo la firmara.

Meia y yo tuvimos una reunión mágica con Elon en nuestra habitación de hotel, donde dio su bendición a los planes detallados para nuestro programa de becas. A Meia le impresionó lo práctico y franco que Elon fue sobre su vida personal, y lo mucho que se interesó por nosotros. Nos preguntó cómo nos habíamos conocido, y le gustó escuchar la detallada historia que contó Meia. Al día siguiente, grabamos una entrevista con él sobre IA segura y por qué quería apoyarla, y todo parecía estar encarrilado.[136]

El momento álgido de la conferencia, el anuncio de la donación de Elon, estaba previsto para las siete de la tarde del domingo 4 de enero de 2015, y yo estaba tan nervioso que la noche anterior la pasé dando vueltas en la cama. Y entonces, apenas quince minutos antes de que nos dirigiésemos a la sesión donde tendría lugar el anuncio, nos topamos con un escollo. El ayudante de Elon llamó para decirnos que parecía que este no podría hacer el anuncio, y Meia dijo que nunca me había visto tan estresado o decepcionado. Finalmente apareció Elon, y mientras hablaba con él pude oír la cuenta atrás previa al inicio de la sesión. Me explicó que estaban a solo dos días del lanzamiento de un cohete crucial de SpaceX, con el que esperaban lograr el primer aterrizaje exitoso de la primera fase del cohete sobre un barco autónomo, y que, como este era un gran hito, el equipo de SpaceX quería evitar las distracciones que conllevaría el hecho de que Elon apareciera en los medios durante esos días. Anthony Aguirre, tranquilo y sensato como siempre, señaló que lo que esto significaba era que nadie quería llamar la atención de los medios, ni Elon ni la comunidad de la IA. Llegamos unos minutos tarde a la sesión en que yo ejercía de moderador, pero teníamos un plan: no se mencionaría ninguna cantidad concreta, para asegurarnos de que el anuncio no tendría interés periodístico, y yo me encargaría de imponer la regla de Chatham House para que todos mantuvieran en secreto el anuncio de

Elon durante nueve días si su cohete llegaba a la estación espacial, tanto si el aterrizaje tenía éxito como si no. Elon nos explicó que necesitaría aún más tiempo si el cohete explotaba durante el lanzamiento.

La cuenta atrás hasta el anuncio llegó finalmente a cero. Los participantes en la mesa redonda sobre superinteligencia que yo había moderado seguían sentados a mi lado en el escenario: Eliezer Yudkowsky, Elon Musk, Nick Bostrom, Richard Mallah, Murray Shanahan, Bart Selman, Shane Legg y Vernor Vinge. La gente fue dejando de aplaudir gradualmente, pero los participantes permanecieron sentados, porque les había dicho que se quedaran ahí sin explicarles por qué. Más tarde, Meia me dijo que en ese momento su pulso estaba disparado, y que agarró la mano tranquilizadora de Viktoriya Krakovna por debajo de la mesa. Sonreí, sabiendo que este era el momento para el que tanto habíamos trabajado y que tanto tiempo llevábamos esperando.

Dije que me alegraba mucho de que hubiera tanto acuerdo en la reunión sobre la necesidad de más investigaciones para que la IA siguiese siendo beneficiosa, y de que hubiera muchas líneas concretas de investigación en las que podríamos trabajar de inmediato. Pero en esa sesión se había hablado de riesgos graves, añadí, por lo que sería agradable levantar el ánimo y animarnos antes de pasar al bar y al banquete que había organizado fuera. «Y por eso le cedo el micrófono a... ¡Elon Musk!» Sentí que estábamos viviendo un momento histórico cuando Elon tomó el micrófono y anunció que donaría una gran cantidad de dinero a la investigación en IA segura. Como era de esperar, se desató el entusiasmo entre los asistentes. Como estaba previsto, Elon no mencionó cuánto dinero donaría, pero yo sabía que eran unos fantásticos diez millones de dólares, como habíamos acordado.

Tras la conferencia, Meia y yo fuimos a visitar a nuestros padres en Suecia y Rumanía, y, conteniendo el aliento, vimos el lanzamiento del cohete en tiempo real con mi padre en Estocolmo. Por desgracia, el intento de aterrizaje terminó con lo que Elon eufemísticamente llama un DRNP, «desmontaje rápido no programado». Su equipo tardaría aún quince meses más en lograr un aterrizaje exitoso sobre un barco.[137] Sin embargo, todos los satélites se pusieron en órbita con éxito, y lo mismo sucedió con nuestro programa de becas, a través de un tuit de Elon a sus millones de seguidores.[138]

POPULARIZAR LA ia SEGURA

Un objetivo clave de la conferencia de Puerto Rico había sido el de popularizar la investigación sobre IA segura, y fue emocionante ver cómo esto ocurría paso a paso. Primero estuvo la conferencia en sí, donde muchos investigadores comenzaron a sentirse cómodos, involucrándose en el tema, una vez que se dieron cuenta de que formaban parte de una creciente comunidad de gente como ellos. Me conmovió profundamente el apoyo que recibí de muchos de los participantes. Por ejemplo, Bart Selman, profesor de IA en la Universidad de Cornell, me envió un correo electrónico diciendo: «Sinceramente, nunca he visto una reunión científica mejor organizada o más emocionante e intelectualmente estimulante que esta».

El siguiente paso hacia la popularización comenzó el 11 de enero, cuando Elon tuiteó: «Los mejores desarrolladores de inteligencia artificial del mundo firman una carta abierta pidiendo que se investigue sobre IA segura»[139], con un enlace a una página de registro que enseguida acumuló más de ocho mil firmas, incluidas las de muchos de los creadores de IA más prominentes del mundo. De pronto se volvió más difícil decir que las personas preocupadas por la IA segura no sabían de qué estaban hablando, porque ahora esto implicaba afirmar que muchos de los principales investigadores en IA no sabían de lo que hablaban. Cuando vimos cómo se hicieron eco de la carta abierta los medios de todo el mundo, nos alegramos de no haber permitido que los periodistas asistieran a la conferencia. Aunque la palabra más alarmista que aparecía en la carta era «escollos», dio pie a titulares como «Elon Musk y Stephen Hawking firman una carta abierta con la esperanza de evitar el motín de los robots», ilustrada con exterminadores asesinos. De los cientos de artículos que vimos, nuestro favorito fue uno que se burlaba de los demás, diciendo que «un titular que evoca visiones de androides esqueléticos pisoteando cráneos humanos convierte una tecnología compleja y transformadora en un espectáculo carnavalesco».[140] Afortunadamente, también había muchos artículos sobrios, que dieron pie a otra dificultad: mantenernos al día con el torrente de nuevas firmas, que debíamos verificar manualmente para preservar nuestra credibilidad y filtrar bromas como «HAL 9000», «Terminator», «Sarah Jeanette Connor» y «Skynet». Para esta y nuestras futuras cartas abiertas, Viktoriya Krakovna y János Krámar

ayudaron a organizar una brigada de revisores voluntarios formada por Jesse Galef, Eric Gastfriend y Revathi Vinoth Kumar trabajando por turnos, de manera que, cuando Revathi se iba a dormir en India, le pasaba el testigo a Eric en Boston, y así sucesivamente.

El tercer paso hacia la popularización comenzó cuatro días después, cuando Elon tuiteó un enlace a nuestro anuncio de que iba a donar diez millones de dólares a la investigación en IA segura.[141] Una semana más tarde, lanzamos un portal web donde investigadores de todo el mundo podían presentarse para competir por esa financiación. Pudimos poner en marcha el sistema de solicitudes tan rápido solo porque Anthony y yo habíamos pasado la década anterior gestionando competiciones similares para subvenciones de física. El proyecto Open Philanthropy, una organización benéfica con sede en California y centrada en donaciones de gran repercusión, se comprometió generosamente a sumar su propia donación al regalo de Elon para que así pudiésemos conceder más becas. No estábamos seguros de cuántos solicitantes tendríamos, ya que el tema era novedoso y el plazo era breve. La respuesta nos sorprendió: cerca de trescientos equipos de todo el mundo solicitaron en total unos cien millones de dólares. Un tribunal formado por profesores de IA y otros investigadores revisó cuidadosamente las propuestas y seleccionó treinta y siete equipos ganadores, que recibieron financiación de hasta tres años. Cuando anunciamos la lista de receptores de las becas, la respuesta de los medios a nuestras actividades fue por primera vez bastante razonable y no incluyó imágenes de robots asesinos. Por fin empezaba a calar la idea de que la IA segura no eran palabras vacías: realmente había trabajo útil por hacer, y muchos grandes equipos de investigación se pusieron manos a la obra.

El cuarto paso en la popularización de la IA segura se produjo de manera natural a lo largo de los dos años siguiente, con decenas de publicaciones técnicas y de talleres sobre la IA segura por todo el mundo, por lo general como parte de las principales conferencias sobre IA. Varias personas tenaces llevaban muchos años intentando involucrar a la comunidad de la IA en la investigación sobre seguridad, con un éxito bastante limitado, pero esta vez la cosa realmente se puso en marcha. Muchas de estas publicaciones recibieron financiación de nuestro programa de becas, y en el FLI hicimos todo lo posible para ayudar a organizar y financiar tantos de estos talleres como pudimos, pero una proporción creciente de los mismos fue posible gracias a

que los investigadores en IA invirtieron en ellos su propio tiempo y recursos. En consecuencia, cada vez más investigadores tuvieron noticia de la investigación en seguridad a través de sus colegas, y descubrieron que, además de ser útil, también podía ser divertida, pues implicaba la resolución de problemas matemáticos y computacionales interesantes.

Por supuesto, no a todo el mundo le parecen divertidas las ecuaciones complicadas. Dos años después de nuestra conferencia de Puerto Rico, como preámbulo de la conferencia de Asilomar, organizamos un seminario técnico donde los ganadores de las becas del FLI pudieron presentar sus investigaciones, y ver en una gran pantalla, diapositiva tras diapositiva, un montón de símbolos matemáticos. Moshe Vardi, profesor de IA en la Universidad de Rice, bromeó diciendo que supo que habíamos conseguido afianzar el campo de la investigación en IA segura cuando las reuniones se volvieron aburridas.

Este crecimiento espectacular del trabajo en IA segura no se notó solo en el mundo académico. Amazon, DeepMind, Facebook, Google, IBM y Microsoft lanzaron un consorcio empresarial dedicado a IA beneficiosa.[142] Nuevas y cuantiosas donaciones destinadas a la investigación en IA segura permitieron intensificar la actividad de las principales organizaciones sin ánimo de lucro similares a la nuestra: el Machine Intelligence Research Institute en Berkeley, el Future of Humanity Institute en Oxford y el Centre for the Study of Existential Risk en Cambridge (Reino Unido). Otras donaciones de diez millones de dólares o más permitieron lanzar otros proyectos relacionados con la IA: el Leverhulme Centre for the Future of Intelligence en Cambridge, el K&L Gates Endowment for Ethics and Computational Technologies en Pittsburgh y el Ethics and Governance of Artificial Intelligence Fund en Miami. Por último, pero no por ello menos importante, con un compromiso de mil millones de dólares, Elon Musk se asoció con otros empresarios para lanzar OpenAI, una compañía sin ánimo de lucro con sede en San Francisco cuyo objetivo es promover la IA beneficiosa. La investigación en IA segura había llegado para quedarse.

Coincidiendo con este crecimiento de la investigación se produjo una oleada de opiniones, tanto individuales como colectivas. El consorcio Partnership on Artificial Intelligence hizo públicos sus principios fundamentales, y el Gobierno estadounidense, la Universidad de Stanford y el IEEE (la organización de profesionales técnicos más grande del mundo)

publicaron sendos extensos informes con listas de recomendaciones, junto con decenas de otros informes y documentos de toma de posición de otras entidades.[143]

Estábamos deseando facilitar una discusión profunda entre los asistentes a Asilomar e identificar en qué cosas (si es que las había) estaba de acuerdo esta comunidad variopinta. Esto llevó a Lucas Perry a asumir la hercúlea tarea de leer todos esos documentos que habían llegado a nuestro conocimiento y hacer un resumen de todas las opiniones que en ellos se expresaban. En un maratoniano esfuerzo iniciado por Anthony Aguirre y concluido con una serie de largas teleconferencias, el equipo del FLI intentó combinar opiniones similares y eliminar la terminología burocrática redundante para acabar obteniendo una única lista de principios sucintos, que también incluía opiniones no publicadas, pero influyentes, que se habían expresado de manera más informal durante las charlas y en otros lugares. Pero esta lista aún contenía muchas ambigüedades y expresiones contradictorias, y dejaba mucho espacio para la interpretación, por lo que el mes anterior a la conferencia la compartimos con los participantes y recopilamos sus opiniones y sugerencias para pulir los principios o introducir otros nuevos. Las contribuciones de la comunidad resultaron en una lista de principios sustancialmente revisada, que sería la que usaríamos en la conferencia.

Ya en Asilomar, la lista se refinó aún más en dos pasos. En primer lugar, grupos pequeños de personas discutieron los principios en los que estaban más interesados (figura 9.4), incorporando matizaciones detalladas, comentarios, nuevos principios y versiones alternativas de los ya existentes. A continuación, hicimos una encuesta entre todos los asistentes para determinar el nivel de apoyo existente para cada versión de cada principio.



FIGURA 9.3. Grupos de grandes cerebros ponderan los principios de IA en Asilomar.

Este proceso colectivo fue exhaustivo y agotador. Anthony, Meia y yo robamos horas al sueño y a las comidas intentando recopilar todo el material necesario a tiempo para los siguientes pasos. Pero también fue emocionante. Después de asistir a discusiones tan minuciosas, espinosas, y en ocasiones polémicas, y de recibir una gran variedad de comentarios, nos sorprendió el alto nivel de consenso que surgió en torno a muchos de los principios durante esa encuesta final (algunos obtuvieron más del 97% de apoyo). Este consenso nos permitió fijar un listón alto para incluir algún principio en la lista final: solo mantuvimos aquellos en los que al menos el 90 % de los asistentes estuvieron de acuerdo. Aunque esto significó que algunos principios famosos se descartaran en el último minuto, incluidos algunos de mis favoritos,[144] permitió que la mayoría de los participantes no tuvieran inconveniente en respaldarlos todos en la hoja de firmas que hicimos circular por el auditorio. Este es el resultado.

LOS PRINCIPIOS DE ASILOMAR PARA LA IA

La inteligencia artificial ya ha proporcionado herramientas beneficiosas que usan a diario personas de todo el mundo. Su desarrollo continuado, guiado por los siguientes principios, brindará extraordinarias oportunidades de ayudar y empoderar a las personas en las décadas y siglos venideros.

Cuestiones relacionadas con la investigación

§1 Objetivo de la investigación: el objetivo en IA no debe ser crear cualquier inteligencia, sino inteligencia beneficiosa.

§2 Financiación de la investigación: la inversión en IA debe ir acompañada de fondos para investigar cómo garantizar que su uso es beneficioso, incluidas cuestiones espinosas relacionadas con la informática, la economía, la legislación, la ética y las ciencias sociales.

a) ¿Cómo podemos hacer que los futuros sistemas de IA sean altamente robustos, que hagan lo que queremos sin que fallen o sean hackeados?

b) ¿Cómo podemos incrementar nuestra prosperidad a través de la automatización sin que las personas pierdan sus recursos ni la sensación de tener un propósito vital?

c) ¿Cómo podemos actualizar nuestros sistemas legales para que sean más justos y eficientes, no queden desfasados respecto a los avances en IA y puedan gestionar los riesgos asociados con la IA?

d) ¿Con qué conjunto de valores habría que conformar la IA, y qué estatus legal y ético debe tener?

§3 Relación entre ciencia y política: debe producirse una comunicación honesta y constructiva entre los investigadores en IA y los responsables políticos.

§4 Cultura de la investigación: debe fomentarse la existencia de una cultura de cooperación, confianza y transparencia entre los investigadores y los desarrolladores de IA.

§5 Evitar las carreras: los equipos que estén desarrollando sistemas de IA deben cooperar activamente para evitar chapuzas en los estándares de seguridad.

Ética y valores

§6 Seguridad: los sistemas de IA deben ser seguros toda su vida operativa y verificables cuando sea posible y factible.

§7 Transparencia en los fallos: si un sistema de IA causa daño, debe ser posible determinar por qué.

§8 Transparencia judicial: cualquier intervención de un sistema autónomo en una decisión debe ir acompañada de una explicación satisfactoria susceptible de ser revisada por una autoridad humana

competente.

§9 Responsabilidad: los desarrolladores de sistemas avanzados de IA no son ajenos a las implicaciones morales del uso, abuso y las acciones de dichos sistemas, pues tienen la responsabilidad y la oportunidad de determinar dichas implicaciones.

§10 Conformidad de valores: los sistemas de IA altamente autónomos deben diseñarse de manera que pueda garantizarse que sus objetivos y comportamientos son conformes con los valores humanos mientras estén en funcionamiento.

§11 Valores humanos: los sistemas de IA deben diseñarse y gestionarse para que sean compatibles con los ideales de dignidad humana, derechos, libertades y diversidad cultural.

§12 Privacidad personal: las personas deben tener el derecho de acceder, gestionar y controlar los datos que generan, dada la capacidad de los sistemas de IA para analizar y utilizar esa información.

§13 Libertad y privacidad: la aplicación de la IA a los datos personales no puede restringir de forma injustificada la libertad, real o percibida, de las personas.

§14 Beneficio compartido: las tecnologías de IA deben beneficiar y empoderar a tanta gente como sea posible.

§15 Prosperidad compartida: la prosperidad económica creada por la IA debe ser ampliamente compartida, para beneficio de toda la humanidad.

§16 Control humano: los seres humanos deben poder decidir si delegan decisiones a los sistemas de IA para lograr objetivos escogidos previamente, y de qué manera lo hacen.

§17 Sin subversión: el poder conferido por el control de sistemas de IA altamente avanzados debe respetar y mejorar los procesos sociales y cívicos de los que depende la salud de la sociedad, y no subvertirlos.

§18 Carrera armamentística: debe evitarse cualquier tipo de carrera armamentística en torno a las armas autónomas letales.

Cuestiones a más largo plazo

§19 Precaución sobre la capacidad: al no haber consenso al respecto, debemos evitar hacer suposiciones fuertes sobre los límites superiores de las capacidades futuras de la IA.

§20 Importancia: la IA avanzada podría representar un profundo cambio en la historia de la vida en la Tierra, y debe planificarse y gestionarse con la atención y los recursos correspondientes.

§21 Riesgos: los riesgos asociados a la IA, especialmente los catastróficos o existenciales, deben estar sujetos a una planificación y unos esfuerzos de mitigación acordes a su impacto potencial.

§22 Automejora recursiva: los sistemas de IA diseñados para automejorarse recursivamente o autorreplicarse de tal forma que pudiera llevar a un rápido aumento cualitativo o cuantitativo deben someterse a unas estrictas medidas de control y seguridad.

§23 Bien común: la superinteligencia solo debe desarrollarse al servicio de unos ideales éticos compartidos y para beneficio de la humanidad, no de un solo Estado u organización.

Una vez que publicamos los principios en la web, la cantidad de firmas creció espectacularmente, y ahora incluye una asombrosa lista de más de mil investigadores en IA y muchos otros pensadores destacados. Si desea añadir su nombre a la lista de signatarios, puede hacerlo aquí:

<http://futureoflife.org/ai-principles>.

Nos llamó la atención no solo el grado de consenso sobre los principios, sino también lo enérgicos que eran. Claro, a primera vista algunos de ellos parecen tan controvertidos como decir que «la paz, el amor y la maternidad son buenos». Pero, en realidad, muchos de ellos tienen garra, como se puede comprobar fácilmente si se formulan en sentido negativo. Por ejemplo, «¡La superinteligencia es imposible!» viola §19, e «¡Investigar cómo reducir el riesgo existencial de la IA es una total pérdida de tiempo!» viola §21.

De hecho, como puede usted comprobar por sí mismo si ve en YouTube nuestra mesa redonda sobre el largo plazo,[145] Elon Musk, Stuart Russell, Ray Kurzweil, Demis Hassabis, Sam Harris, Nick Bostrom, David Chalmers, Bart Selman y Jaan Tallinn coincidieron en que probablemente se desarrollará la superinteligencia y en que la investigación en seguridad es importante.

Confío en que los principios de Asilomar para IA servirán de punto de partida para más reflexiones incisivas, que en última instancia conducirán a una sensibilización de las estrategias y políticas de la IA. Con esta intención, nuestro director de comunicaciones de FLI, Ariel Conn, trabajó con Tucker Davey y otros miembros del equipo para entrevistarse con investigadores punteros en IA sobre estos principios y cómo los interpretaban ellos, mientras David Stanley y su equipo internacional de voluntarios de FLI traducían estos principios en las principales lenguas el mundo.

OPTIMISMO CONSCIENTE

Como confesé al principio de este epílogo, me siento más optimista sobre el futuro de la vida de lo que me he sentido en mucho tiempo. Para explicar por qué, conté mi historia personal.

Mis experiencias en los últimos años han hecho que aumente mi optimismo por dos razones diferentes. En primer lugar, he sido testigo de cómo la comunidad de la IA aúna esfuerzos de manera extraordinaria para abordar constructivamente los desafíos futuros, a menudo en colaboración con pensadores de otros campos. Después de la conferencia de Asilomar, Elon me dijo que le sorprendía que la seguridad de la IA hubiera pasado de ser un tema marginal a convertirse en un asunto central en solo unos pocos

años, y yo estoy tan sorprendido como él. Y no son solo las cuestiones a corto plazo del capítulo 3 las que están siendo objeto de debate, sino incluso la superinteligencia y el riesgo existencial, como es patente en los principios de Asilomar para la IA. No habría habido manera de aprobar esos principios en Puerto Rico dos años antes, cuando la palabra más inquietante que apareció en la versión final de la carta abierta era «escollos».

Me gusta observar a la gente, y, durante la última mañana de la conferencia de Asilomar, permanecí un momento en un lateral del auditorio y me fijé en los participantes que escuchaban un debate sobre IA y legislación. Para mi sorpresa, me embargó una difusa sensación de calidez, y de pronto me sentí muy conmovido. ¡Esto parecía tan distinto de Puerto Rico! Por aquel entonces, recuerdo que veía a la mayoría de la comunidad de IA con una mezcla de miedo y respeto: no exactamente como un equipo rival, sino como un grupo al que mis colegas preocupados por la IA y yo sentíamos que debíamos convencer. Pero ahora parecía evidente que todos estábamos en el mismo equipo. Como probablemente haya usted comprendido al leer este libro, todavía no sé cómo crear un gran futuro con la IA, por lo que me parece maravilloso formar parte de una comunidad en crecimiento que busca unida las respuestas.



FIGURA 9.4. Una comunidad creciente busca conjuntamente respuestas en Asilomar.

La segunda razón por la que me he vuelto más optimista es porque la experiencia del FLI ha sido enriquecedora. Lo que había provocado mis lágrimas en Londres fue una sensación de inevitabilidad: que quizá teníamos ante nosotros un futuro inquietante y no pudiéramos hacer nada al respecto. Pero los tres años siguientes disiparon mi pesimismo fatalista. Si incluso un variopinto grupo de voluntarios no remunerados puede ejercer una influencia positiva en el que considero que es el debate más importante de nuestro tiempo, imaginemos lo que podemos hacer si todos trabajamos juntos.

En la charla que dio en Asilomar, Erik Brynjolfsson habló de dos tipos de optimismo. El primero es el incondicional, como la expectativa segura de que mañana por la mañana saldrá el sol. Luego está lo que llamó el «optimismo consciente», que es la expectativa de que sucederán cosas buenas si hacemos planes y nos esforzamos para llevarlos a la práctica. Ese es el tipo de optimismo que ahora me inspira el futuro de la vida.

¿Qué puede usted hacer para tener una influencia positiva sobre el futuro de la vida, ahora que entramos en la era de la IA? Por razones que explicaré enseguida, creo que un gran primer paso es trabajar para convertirse en un optimista consciente, si no lo es ya. Para conseguirlo, es fundamental tener una visión positiva del futuro. Cuando los estudiantes del MIT vienen a mi oficina en busca de orientación profesional, normalmente comienzo preguntándoles dónde se ven en una década. Si una estudiante respondiera: «Quizá estaré en el hospital con cáncer, o en un cementerio tras haber sido atropellada por un autobús», me pondría firme con ella. ¡Visualizar solo los futuros negativos es una estrategia pésima para planificar la carrera profesional! Dedicar todos nuestros esfuerzos a evitar enfermedades y accidentes es una manera casi segura de acabar hipocondriacos y paranoicos, no de conseguir ser felices. Lo que me gustaría es escucharla describir sus objetivos con entusiasmo, tras lo cual podríamos discutir las estrategias para alcanzarlos evitando todos los escollos.

Erik señaló que, según la teoría de juegos, las visiones positivas forman la base de buena parte de toda la colaboración que se da en el mundo, desde los matrimonios y las fusiones corporativas hasta la decisión de varios estados independientes de formar los Estados Unidos. Al fin y al cabo, ¿por qué sacrificar algo que tenemos si no podemos imaginar la ganancia aún mayor

que obtendremos al hacerlo? Esto significa que debemos imaginarnos un futuro positivo no solo para nosotros mismos, sino también para la sociedad y para la humanidad entera. En otras palabras, ¡necesitamos más esperanza existencial! Sin embargo, como a Meia le gusta recordarme, desde Frankenstein hasta Terminator, las visiones futuristas en la literatura y en el cine son predominantemente distópicas. Dicho de otro modo, nosotros, como sociedad, estamos planificando nuestro futuro tan mal como esa estudiante del MIT. Por eso necesitamos más optimistas conscientes. Y esta es la razón por la que a lo largo de este libro he intentado que usted pensase en qué tipo de futuro *desea* y no simplemente en qué tipo de futuro *teme*, para que encuentre metas compartidas en torno a las cuales planificar y trabajar.

A lo largo de este libro, hemos visto cómo es probable que la IA nos brinde grandes oportunidades y también presente desafíos difíciles. Una estrategia que posiblemente sea útil para hacer frente a todos los desafíos de la IA es que actuemos juntos para mejorar la sociedad humana antes de que la IA despegue por completo. Nos interesa educar a nuestros jóvenes para que hagan que la tecnología sea robusta y beneficiosa antes de cederle mucho poder. Nos conviene modernizar las leyes antes de que la tecnología las vuelva obsoletas. Nos interesa resolver los conflictos internacionales antes de que desencadenen una carrera armamentista de armas autónomas. Somos responsables de crear una economía que asegure la prosperidad a todos antes de que la IA amplifique las desigualdades. Nos interesa tener una sociedad en la que los resultados de la investigación en IA segura se implementan en lugar de ignorarse. Y mirando más lejos hacia el futuro, a los desafíos relacionados con la IAG sobrehumana, nos conviene acordar al menos unos estándares éticos básicos antes de comenzar a inculcar estos estándares a máquinas potentes. En un mundo polarizado y caótico, quienes tengan poder para usar la IA con propósitos maliciosos tendrán más motivación y capacidad para hacerlo, y los equipos que compiten para desarrollar IAG sentirán más presión para relajar la seguridad que para cooperar. En resumen, si podemos crear una sociedad humana más armoniosa, caracterizada por la cooperación en torno a objetivos compartidos, esto mejorará las perspectivas de que la revolución de la IA tenga un final feliz.

En otras palabras, una de las maneras más efectivas de mejorar el futuro de la vida es mejorar el mañana. Tenemos la capacidad de hacerlo de muchas formas. Por supuesto, podemos votar en las urnas y decir a los políticos lo

que pensamos sobre la educación, la privacidad, las armas letales autónomas, el desempleo tecnológico y otros asuntos. Pero también votamos todos los días a través de lo que decidimos comprar, las noticias que decidimos consumir, lo que decidimos compartir y el ejemplo que damos con nuestro comportamiento. ¿Queremos ser alguien que interrumpe todas sus conversaciones para mirar su móvil, o alguien que se sienta capaz de usar la tecnología de forma planificada y consciente? ¿Queremos ser dueños de la tecnología o que la tecnología se adueñe de nosotros? ¿Qué queremos que signifique ser humano en la era de la IA? Por favor, hable de todo esto con las personas que tiene a su alrededor: esta conversación no solo es importante, sino también fascinante.

Somos los custodios del futuro de la vida, ahora que estamos configurando cómo será la era de la IA. Aunque lloré en Londres, ahora siento que este futuro no tiene nada de inevitable, y sé que es mucho más fácil de lo que pensaba tener una influencia positiva. El futuro no está decidido ni grabado en piedra; somos nosotros quienes lo creamos. ¡Creemos juntos un futuro que nos ilusione!

AGRADECIMIENTOS

Estoy sinceramente agradecido a todos aquellos que me han alentado y ayudado a escribir este libro, entre quienes están:

Mi familia, amigos, profesores, colegas y colaboradores, por su apoyo e inspiración a lo largo de los años;

mi madre, por fomentar mi curiosidad por la consciencia y el sentido de las cosas;

mi padre, por su espíritu luchador para hacer del mundo un lugar mejor;

mis hijos, Philip y Alexander, por mostrarme lo asombroso que es ver surgir inteligencia de nivel humano;

todos los entusiastas de la ciencia y la tecnología de todo el mundo que se han puesto en contacto conmigo a lo largo de los años con preguntas, comentarios y expresiones de apoyo para explorar y publicar mis ideas;

mi agente, John Brockman, por insistir hasta que accedí a escribir este libro;

Bob Penna, Jesse Thaler y Jeremy England, por fructíferas discusiones sobre cuásares, esfalerones y termodinámica, respectivamente;

las personas que me dieron su opinión sobre partes del manuscrito, como mi madre, mi hermano Per, Luisa Bahet, Rob Bensinger, Katerina Bergström, Erik Brynjolfsson, Daniela Chita, David Chalmers, Nima Deghani, Henry Lin, Elin Malmsköld, Toby Ord, Jeremy Owen, Lucas Perry, Anthony Romero, Nate Soares y Jaan Tallinn;

los superhéroes que me hicieron comentarios sobre borradores del libro entero, a saber, Meia, mi padre, Anthony Aguirre, Paul Almond, Matthew Graves, Phillip Helbig, Richard Mallah, David Marble, Howard Messing, Luiño Seoane, Marin Soljačić, mi editor Dan Frank y, por encima de todos;

Meia, mi amada musa y compañera de viaje, por su aliento, apoyo e inspiración eternos, sin los que este libro no existiría.

ÍNDICE ALFABÉTICO

NOTAS

BIENVENIDOS A LA CONVERSACIÓN MÁS IMPORTANTE DE NUESTRO TIEMPO

[1] «The AI Revolution: Our Immortality or Extinction?», *Wait But Why* (27 de enero de 2015), en <https://waitbutwhy.com/2015/01/artificial-intelligence-revolution-2.html>.
[2] Esta carta abierta, «Research Priorities for Robust and Beneficial Artificial Intelligence», puede encontrarse en <https://futureoflife.org/ai-open-letter/>.
[3] Ejemplo del alarmismo en torno a los robots habitual en los medios de comunicación: Ellie Zolfagharifard, «Artificial Intelligence "Could Be the Worst Thing to Happen to Humanity"», *Daily Mail*, 2 de mayo de 2014, en <http://tinyurl.com/hawkingbots>.

LA MATERIA SE VUELVE INTELIGENTE

[4] Notas sobre el origen de la expresión IAG, en <http://goertzel.org/who-coined-the-term-agi/>.
[5] Hans Moravec, «When Will Computer Hardware Match the Human Brain?», *Journal of Evolution and Technology* (1998), vol. 1.
[6] En la figura que muestra la potencia de computación respecto a los años, los datos anteriores a 2011 proceden del libro de Ray Kurzweil *Cómo crear una mente*, y los posteriores se han computado a partir de las referencias en <https://en.wikipedia.org/wiki/FLOPS>.
[7] El pionero de la computación cuántica David Deutsch describe cómo entiende que la computación cuántica es evidencia de la existencia de universos paralelos en su libro *The Fabric of Reality: The Science of Parallel Universes—and Its Implications* (Londres, Allen Lane, 1997) [Hay trad. cast.: *La estructura de la realidad*, Barcelona, Anagrama, 2002]. Si quiere conocer mis propias ideas sobre los universos paralelos cuánticos como el tercero de cuatro niveles del multiverso, las encontrará en mi libro anterior *Our Mathematical Universe: My Quest for the Ultimate Nature of Reality* (Nueva York, Knopf, 2014) [Hay trad. cast.: *Nuestro universo matemático*, Barcelona, Antoni Bosch, 2017].
[8] Entrada de Google en la IA de reconocimiento de imágenes, en < https://arxiv.org/pdf/1411.4555.pdf>.

EL FUTURO PRÓXIMO

[9] Véase «Google DeepMind's Deep Q-learning Playing Atari Breakout», en YouTube en <https://tinyurl.com/atariai>.
[10] Véase Volodymyr Mnih *et al.*, «Human-Level Control Through Deep Reinforcement Learning», *Nature* 518, 26 de febrero de 2015, pp. 529-533. Disponible online en <http://tinyurl.com/ataripaper>.
[11] Aquí puede verse un vídeo del robot Big Dog en acción, en <https://www.youtube.com/watch?

v=W1czBcnX1Ww>.
[12] Para las reacciones al movimiento sorprendentemente creativo por parte de AlphaGo en la línea 5, véase «Move 37!! Lee Sedol vs AlphaGo Match 2», en <https://www.youtube.com/watch?v=JNrXgpSEEIE>.
[13] Demis Hassabis describe las reacciones a AlphaGo de jugadores humanos de go, en <https://www.youtube.com/watch?v=otJKzpNWZT4>.
[14] Para las mejoras recientes que ha experimentado la traducción automática, véase Gideon Lewis-Kraus, «The Great A.I. Awakening», *New York Times Magazine* (14 de diciembre de 2016), accesible online en <http://www.nytimes.com/2016/12/14/magazine/the-great-ai-awakening.html>. GoogleTranslate está disponible en <https://translate.google.com>.
[15] Competición en torno al desafío de los esquemas de Winograd, en <http://tinyurl.com/winogradchallenge>.
[16] Vídeo de la explosión del Ariane 5, en <https://www.youtube.com/watch?v=qnHn8W1Em6E>.
[17] Informe sobre el fracaso del vuelo 501 del Ariane 5 elaborado por la comisión de investigación, en <http://tinyurl.com/arianeflop>.
[18] Informe de la comisión de investigación sobre la fase 1 del accidente del Mars Climate Orbiter de la NASA <http://tinyurl.com/marsflop>.
[19] Según la información más detallada y consistente de que disponemos, lo que provocó el fallo de la misión Mariner 1 a Venus fue la incorrecta transcripción a mano de un solo símbolo matemático (la falta de una barra superior), en <http://tinyurl.com/marinerflop>.
[20] Una descripción detallada del fallo de la misión soviética Fobos 1 a Marte se puede encontrar en Wesley T. Huntress Jr. y Mikhail Ya. Marov, *Soviet Robots in the Solar System*, Nueva York, Praxis Publishing, 2011, p. 308.
[21] Cómo un software sin verificar le costó a Knight Capital 440 millones de dólares en 45 minutos, en <http://tinyurl.com/knightflop1> y <http://tinyurl.com/knightflop2>.
[22] Informe del Gobierno estadounidense sobre la «quiebra instantánea» de Wall Street: «Conclusiones sobre lo acaecido en el mercado el 6 de mayo 2010» (30 de septiembre de 2010), en <http://tinyurl.com/flashcrashreport>.
[23] Impresión 3D de edificios (<https://www.youtube.com/watch?v=SObzNdyRTBs>), dispositivos micromecánicos (<http://tinyurl.com/tinyprinter>) y muchas cosas entremedias (<https://www.youtube.com/watch?v=xVU4FLrsPXs>).
[24] Mapa mundial de los *fab labs* comunitarios, en <https://www.fablabs.io/labs/map>.
[25] Artículo de prensa sobre la muerte de Robert Williams a manos de un robot industrial, en <http://tinyurl.com/williamsaccident>.
[26] Artículo de prensa sobre la muerte de Kenji Urada a manos de un robot industrial, en <http://tinyurl.com/uradaaccident>.
[27] Artículo de prensa sobre la muerte del trabajador de Volkswagen a manos de un robot industrial: <http://tinyurl.com/baunatalaccident>.
[28] Informe del Gobierno estadounidense sobre muertes laborales, en <https://www.osha.gov/dep/fatcat/dep_fatcat.html>.
[29] Estadísticas de víctimas de accidentes de circulación, en <http://tinyurl.com/roadsafety2> y en <http://tinyurl.com/roadsafety3>.
[30] Sobre la primera víctima mortal del piloto automático de Tesla, véase Andrew Buncombe, «Tesla Crash: Driver Who Died While on Autopilot Mode "Was Watching Harry Potter"», *Independent* (1 de julio de 2016), en <http://tinyurl.com/teslacrashstory>. Para el informe de la Office of Defects Investigation de la National Highway Traffic Safety Administration estadounidense, véase <http://tinyurl.com/teslacrashreport>.
[31] Para más información sobre el desastre del *Herald of Free Enterprise*, véase R. B. Whittingham,

The Blame Machine: Why Human Error Causes Accidents, Oxford (Reino Unido), Elsevier, 2004.

[32] Documental sobre el accidente del vuelo 447 de Air France, en <https://www.youtube.com/watch?v=dpPkp8OGQFI>; informe del accidente: <http://tinyurl.com/af447report>; análisis externo: <http://tinyurl.com/thomsonarticle>.

[33] Informe oficial sobre el apagón de 2003 en Estados Unidos y Canadá, en <http://tinyurl.com/uscanadablackout>.

[34] Informe final de la comisión presidencial sobre el accidente de Three Mile Island, en <http://www.threemileisland.org/downloads/188.pdf>.

[35] Estudio holandés que demuestra cómo la IA puede competir con los radiólogos humanos en el diagnóstico de cáncer de próstata a partir de MRI, en <http://tinyurl.com/prostate-ai>.

[36] Estudio de Stanford que demuestra que la IA puede diagnosticar el cáncer de pulmón mejor que los patólogos humanos, en <http://tinyurl.com/lungcancer-ai>.

[37] Investigación de los accidentes con la máquina de radioterapia Therac-25, en <http://tinyurl.com/theracfailure>.

[38] Informe sobre las sobredosis letales de radiación en Panamá provocadas por una interfaz de usuario confusa, en <http://tinyurl.com/cobalt60accident>.

[39] Estudio sobre incidentes adversos en cirugía robotizada, en <https://arxiv.org/abs/1507.03518>.

[40] Artículo sobre el número de muertes debidas a la deficiente atención hospitalaria, en <http://tinyurl.com/medaccidents>.

[41] Yahoo estableció un nuevo estándar de lo que se entiende por un «gran hackeo» cuando anunció que habían penetrado en mil millones de sus cuentas de su usuario: <https://www.wired.com/2016/12/yahoo-hack-billion-users/>.

[42] Artículo de *The New York Times* sobre la absolución y posterior condena del asesino del KKK, en <http://tinyurl.com/kkkacquittal>.

[43] El estudio de 2011 de Danziger *et al.* (en <http://www.pnas.org/content/108/17/6889.full>), que argumenta que los jueces hambrientos son más severos, fue tildado de erróneo por Keren Weinshall-Margela y John Shapard (en <http://www.pnas.org/content/108/42/E833.full>), pero Danziger *et al.* insisten en que sus afirmaciones siguen siendo válidas (en <http://www.pnas.org/content/108/42/E834.full>).

[44] Informe de *Pro Publica* sobre la existencia de sesgo racial en el software de predicción de la reincidencia, en <http://tinyurl.com/robojudge>.

[45] El uso de IRMf y otras técnicas de escaneo cerebral como evidencia en un juicio es una cuestión muy polémica, como también lo es la fiabilidad de dichas técnicas, aunque muchos grupos afirman obtener resultados positivos superiores al 90 %, en <http://journal.frontiersin.org/article/10.3389/fpsyg.2015.00709/full>.

[46] PBS grabó una película sobre el incidente, titulada *The Man Who Saved the World*, en la cual Vasili Arjípov por sí solo evitaba un ataque nuclear soviético, en <https://www.youtube.com/watch?v=4VPY2SgyG5w>.

[47] La historia de cómo Stanislav Petrov hizo caso omiso de los avisos de un ataque nuclear estadounidense y consideró que se trataba de una falsa alarma se llevó al cine con el título *The Man Who Saved the World* (no confundir con la película del mismo título que se menciona en la nota anterior), y Petrov fue homenajeado en Naciones Unidas y recibió el Premio al Ciudadano del Mundo, en <https://www.youtube.com/watch?v=IncSjwWQHMo>.

[48] Carta abierta sobre armas autónomas de investigadores en IA y robótica, en <http://futureoflife.org/open-letter-autonomous-weapons/>.

[49] Este funcionario estadounidense aparentemente desea una carrera armamentística en IA militar, en <http://tinyurl.com/workquote>.

[50] Estudio de la desigualdad económica en Estados Unidos desde 1913, en <http://gabriel-

zucman.eu/files/SaezZucman2015.pdf>.
[51] Informe de Oxfam sobre la desigualdad económica mundial, en <http://tinyurl.com/oxfam2017>.
[52] Para una excelente introducción a la hipótesis de que la desigualdad es debida a la tecnología, véase Erik Brynjolfsson y Andrew McAfee, *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies* (Nueva York, Norton, 2014) [Hay trad. cast.: *La segunda era de las máquinas. Trabajo, progreso y prosperidad en una era de brillantes tecnologías*, Buenos Aires, Temas Grupo Editorial, 2013].
[53] Artículo en *The Atlantic* sobre la caída de los salarios para las personas con menor formación, en <http://tinyurl.com/wagedrop>.
[54] Los datos que se representan están tomados de Facundo Alvaredo, Anthony B. Atkinson, Thomas Piketty, Emmanuel Saez y Gabriel Zucman, *The World Wealth and Income Database* (<http://www.wid.world>), incluidas las ganancias de capital.
[55] Presentación de James Manyika que muestra cómo la renta se desplaza de la mano de obra al capital, en <http://futureoflife.org/data/PDF/james_manyika.pdf>.
[56] Pronósticos sobre la automatización de distintos trabajos en el futuro según Oxford University (en <http://tinyurl.com/automationoxford>) y McKinsey (en <http://tinyurl.com/automationmckinsey>).
[57] Vídeo de un cocinero robotizado, en <https://www.youtube.com/watch?v=fE6i2OO6Y6s>.
[58] Marin Soljačić analizó estas opciones en el seminario Computers Gone Wild: Impact and Implications of Developments in Artificial Intelligence on Society, celebrado en 2016, en <http://futureoflife.org/2016/05/06/computers-gone-wild/>.
[59] Las propuestas de Andrew McAfee para crear más puestos de trabajo buenos, en <http://futureoflife.org/data/PDF/andrew_mcafee.pdf>.
[60] Además de los muchos artículos académicos que argumentan que «esta vez es diferente» para el desempleo tecnológico, el vídeo «Humans Need Not Apply» expone de forma concisa esa misma idea: <https://www.youtube.com/watch?v=7Pq-S557XQU>.
[61] Oficina de Estadísticas Laborales estadounidense, en <http://www.bls.gov/cps/cpsaat11.htm>.
[62] El argumento de que «esta vez es diferente» para el desempleo tecnológico: Federico Pistono, *Robots Will Steal Your Job, but That's OK* (2012), en <http://robotswillstealyourjob.com>.
[63] Variaciones en el número de caballos en Estados Unidos, en <http://tinyurl.com/horsedecline>.
[64] Metaanálisis que demuestra cómo el desempleo afecta al bienestar: Maike Luhmann *et al.*, «Subjective Well-Being and Adaptation to Life Events: A Meta-Analysis», *Journal of Personality and Social Psychology* 102, n.º 3 (2012), p. 592; disponible online en <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3289759>.
[65] Estudios de qué cosas hacen que aumente la sensación de bienestar de las personas: Angela Duckworth, Tracy Steen y Martin Seligman, «Positive Psychology in Clinical Practice», *Annual Review of Clinical Psychology* 1 (2005), pp. 629-651, disponible online en <http://tinyurl.com/wellbeingduckworth>. Weiting Ng y Ed Diener, «What Matters to the Rich and the Poor? Subjective Well-Being, Financial Satisfaction, and Postmaterialist Needs Across the World», *Journal of Personality and Social Psychology* 107, n.º 2 (2014), p. 326, disponible online en <http://psycnet.apa.org/journals/psp/107/2/326>. Kirsten Weir, «More than Job Satisfaction», *Monitor on Psychology* 44, n.º 11 (diciembre de 2013), online en <http://www.apa.org/monitor/2013/12/job-satisfaction.aspx>.
[66] Multiplicar entre sí alrededor de 10^{11} neuronas, unas 10^{4} conexiones por neurona y en torno a una (10^{0}) activación por neurona cada segundo podría dar a entender que alrededor de 10^{15} FLOPS (un petaFLOPS) basta para simular un cerebro humano, pero existen muchas complicaciones que aún apenas se conocen, incluida la secuencia temporal detallada de las activaciones y la cuestión de si también deberían simularse partes pequeñas de las neuronas y sinapsis. Dharmendra Modha,

informático de IBM, ha estimado que se requieren 38 petaFLOPS (en <http://tinyurl.com/javln43>), mientras que el neurocientífico Henry Markram calcula que se necesitan alrededor de mil petaFLOPS (en <http://tinyurl.com/6rpohqv>). Katja Grace y Paul Christiano, investigadores en IA, argumentan que el aspecto más costoso de la simulación del cerebro no es la computación sino la comunicación, y esa es también una tarea del orden de magnitud de lo que los ordenadores actuales son capaces de hacer, en <http://aiimpacts.org/about>.
[67] Para una estimación interesante de la capacidad de computación del cerebro humano, véase Hans Moravec, «When Will Computer Hardware Match the Human Brain?», *Journal of Evolution and Technology*, (1998), vol. 1.

¿EXPLOSIÓN DE INTELIGENCIA?

[68] Para un vídeo del primer pájaro mecánico, véase Markus Fischer, «Un robot que vuela como un pájaro», charla TED, julio de 2011, en <https://www.ted.com/talks/a_robot_that_flies_like_a_bird?language=es>.

TRAS LA EXPLOSIÓN

[69] Ray Kurzweil, *The Singularity Is Near* (Nueva York, Viking Press, 2005) [Hay trad. cast.: *La singularidad está cerca*, Berlín, Lola Books, 2015].
[70] El escenario de la «IA niñera» de Ben Goertzel se describe en <https://wiki.lesswrong.com/wiki/Nanny_AI>.
[71] Para una discusión sobre la relación entre máquinas y humanos, y sobre si las máquinas son nuestros esclavos, véase Benjamin Wallace-Wells, «Boyhood», *New York magazine* (20 de mayo de 2015), disponible online en <http://tinyurl.com/aislaves>.
[72] El crimen contra la mente se discute en el libro de Nick Bostrom *Superinteligencia*, y, con una mayor profundidad técnica, en este artículo científico reciente: Nick Bostrom, Allan Dafoe y Carrick Flynn, «Policy Desiderata in the Development of Machine Superintelligence» (2016), en <http://www.nickbostrom.com/papers/aipolicy.pdf>.
[73] Matthew Schofield, «Memories of Stasi Color Germans' View of U.S. Surveillance Programs», *McClatchy DC Bureau* (26 de junio de 2013), disponible online en <http://www.mcclatchydc.com/news/nation-world/national/article24750439.html>.
[74] Para leer reflexiones estimulantes sobre cómo los incentivos que tienen las personas pueden llevar a situaciones que nadie desea, recomiendo «Meditations on Moloch», en <http://slatestarcodex.com/2014/07/30/meditations-on-moloch>.
[75] Para una cronología interactiva de los incidentes que podían haber desencadenado accidentalmente una guerra nuclear, véase Future of Life Institute, «Accidental Nuclear War: A Timeline of Close Calls», en <http://tinyurl.com/nukeoops>.
[76] Para las indemnizaciones pagadas a las víctimas de las pruebas nucleares estadounidenses, véase el sitio web del departamento de Justicia estadounidense, «Awards to Date 4/24/2015», en <https://www.justice.gov/civil/awards-date-04242015>.
[77] *Report of the Commission to Assess the Threat to the United States from Electromagnetic Pulse (EMP) Attack*, abril de 2008, disponible online en <http://www.empcommission.org/docs/A2473-EMP_Commission-7MB.pdf>.

[78] Investigaciones independientes realizadas por científicos estadounidenses y soviéticos alertaron a Reagan y a Gorbachov del riesgo de un invierno nuclear: P. J. Crutzen y J. W. Birks, «The Atmosphere After a Nuclear War: Twilight at Noon», *Ambio* 11, n.º 2/3 (1982), pp. 114-125. R. P. Turco, O. B. Toon, T. P. Ackerman, J. B. Pollack y C. Sagan, «Nuclear Winter: Global Consequences of Multiple Nuclear Explosions», *Science* 222 (1983), pp. 1283-1292. V. V. Aleksandrov y G. L. Stenchikov, «On the Modeling of the Climatic Consequences of the Nuclear War», *Proceeding on Applied Mathematics* (Moscú, Centro de Computación de la Academia de Ciencias de la URSS, 1983), p. 21. A. Robock, «Snow and Ice Feedbacks Prolong Effects of Nuclear Winter», *Nature* 310 (1984), pp. 667-670.
[79] Cálculo de los efectos sobre el clima de una guerra nuclear global: A. Robock, L. Oman y L. Stenchikov, «Nuclear Winter Revisited with a Modern Climate Model and Current Nuclear Arsenals: Still Catastrophic Consequences», *Journal of Geophysical Research* 12 (2007), D13107.

NUESTRA HERENCIA CÓSMICA

[80] Para más información, véase Anders Sandberg, «Dyson Sphere FAQ», en <http://www.aleph.se/nada/dysonFAQ.html>.
[81] El artículo seminal de Freeman Dyson sobre sus esferas epónimas: Freeman Dyson, «Search for Artificial Stellar Sources of Infrared Radiation», *Science*, vol. 131 (1959), pp. 1667-1668.
[82] Louis Crane y Shawn Westmoreland explican su propuesta de motor impulsado por un agujero negro en «Are Black Hole Starships Possible?», disponible online en <http://arxiv.org/pdf/0908.1803.pdf>.
[83] Para una buena infografía del CERN que resume las partículas elementales conocidas, véase <http://tinyurl.com/cernparticle>.
[84] Este excepcional vídeo de un prototipo no nuclear del Orión ilustra la idea de un cohete propulsado por bombas nucleares, en <https://www.youtube.com/watch?v=E3Lxx2VAYi8>.
[85] Aquí puede leerse una introducción instructiva a la propulsión láser a vela: Robert L. Forward, «Roundtrip Interstellar Travel Using Laser-Pushed Lightsails», *Journal of Spacecraft and Rockets* 21, n.º 2 (marzo-abril 1984), disponible online en <http://www.lunarsail.com/LightSail/rit-1.pdf>.
[86] Jay Olson analiza la expansión cósmica de las civilizaciones en «Homogeneous Cosmology with Aggressively Expanding Civilizations», *Classical and Quantum Gravity* 32 (2015), disponible online en <http://arxiv.org/abs/1411.4359>.
[87] El primer análisis científico riguroso de nuestro futuro remoto: Freeman J. Dyson, «Time Without End: Physics and Biology in an Open Universe», *Reviews of Modern Physics* 51, n.º 3 (1979), p. 447, disponible online en <http://blog.regehr.org/extra_files/dyson.pdf>.
[88] La fórmula mencionada antes de Seth Lloyd nos dice que realizar una operación de computación durante un intervalo de tiempo cuesta una energía $E \geq \hbar/4\tau$, donde $\hbar$ es la constante de Planck. Si queremos realizar N operaciones una a continuación de la otra (en serie) durante un tiempo T, entonces $\tau = T/N$, de forma que $E/N \geq \hbar N/4T$, lo que nos dice que podemos realizar $N \leq 2\sqrt{ET/\hbar}$ operaciones en serie usando una energía E en un tiempo T. Así pues, tanto la energía como el tiempo son recursos que interesa tener en abundancia. Si repartimos nuestra energía entre n computaciones distintas en paralelo, pueden ejecutarse de forma más lenta y eficiente, dando $N \leq 2\sqrt{ETn/\hbar}$. Nick Bostrom estima que simular una vida humana de cien años requiere en torno *a* N = ^{10}27 operaciones.
[89] Si quiere escuchar un argumento cuidadoso que explica por qué para que surja la vida podría ser necesaria una casualidad sumamente improbable, lo que situaría a nuestros vecinos más cercanos a más de $^{101.0}$00 metros de distancia, le recomiendo este vídeo de Edwin Turner, físico y astrobiólogo en

Princeton: «Improbable Life: An Unappealing but Plausible Scenario for Life’s Origin on Earth», en <https://www.youtube.com/watch?v=Bt6n6Tulbeg>.
[90] Ensayo de Martin Rees sobre la búsqueda de inteligencia extraterrestre, en <https://www.edge.org/annual-question/2016/response/26665>.

OBJETIVOS

[91] Una discusión accesible del trabajo de Jeremy England sobre «adaptación motivada por la disipación» se puede encontrar en Natalie Wolchover, «A New Physics Theory of Life», *Scientific American* (28 de enero de 2014), disponible online en <https://www.scientificamerican.com/article/a-new-physics-theory-of-life/>. *Order Out of Chaos: Man’s New Dialogue with Nature* (Nueva York, Bantam, 1984), de Ilya Prigogine e Isabelle Stengers, sienta buena parte de las bases de este trabajo.
[92] Para más información sobre los sentimientos y sus raíces fisiológicas: William James, *Principles of Psychology* (Nueva York, Henry Holt & Co., 1890); Robert Ornstein, *Evolution of Consciousness: The Origins of the Way We Think* (Nueva York, Simon & Schuster, 1992); António Damásio, *Descartes’ Error: Emotion, Reason, and the Human Brain* (Nueva York, Penguin, 2005) [hay trad. cast.: *El error de Descartes. La emoción, la razón y el cerebro humano*, Barcelona, Crítica, 2010]; y António Damásio, *Self Comes to Mind: Constructing the Conscious Brain* (Nueva York, Vintage, 2012) [hay trad. cast.: *Y el cerebro creó al hombre: ¿Cómo pudo el cerebro generar emociones, sentimientos, ideas y el yo*?, Barcelona, Crítica, 2010].
[93] Eliezer Yudkowsky ha reflexionado sobre cómo conformar los objetivos de la IA amigable no con nuestros objetivos actuales, sino con nuestra voluntad coherente extrapolada (CEV, por sus siglas en inglés). A grandes rasgos, la CEV se define como lo que una versión idealizada de nosotros querría si supiésemos más, pensásemos más rápido y nos pareciésemos más a las personas que querríamos ser. Yudkowsky empezó a criticar el concepto de CEV poco después de proponerlo en 2004 (en <http://intelligence.org/files/CEV.pdf>), tanto por ser demasiado difícil de implementar como porque no está claro que pueda resultar en algo bien definido.
[94] En el enfoque del aprendizaje por refuerzo inverso, una idea esencial es que la IA no está intentado alcanzar sus propios objetivos sino los de su propietario humano. Por lo tanto, tiene motivos para ser precavida cuando alberga dudas sobre lo que su dueño quiere, y para esforzarse por averiguarlo. Tampoco debería tener inconveniente en que su dueño la apagara, puesto que esto significaría que había interpretado erróneamente lo que este quería realmente.
[95] El artículo de Steve Omohundro sobre la emergencia de objetivos en la IA, «The Basic AI Drives», puede encontrarse en <http://tinyurl.com/omohundro2008>. Se publicó originalmente en *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, Pei Wang, Ben Goertzel y Stan Franklin, eds. (Ámsterdam, IOS, 2008), pp. 483-492.
[96] Un libro estimulante y controvertido sobre lo que ocurre cuando la inteligencia se aplica a cumplir órdenes ciegamente sin cuestionar su fundamento moral: Hannah Arendt, *Eichmann in Jerusalem: A Report on the Banality of Evil* (Nueva York, Penguin, 1963) [hay trad. cast.: *Eichmann en Jerusalén*, Barcelona, Lumen, 2013]. Un dilema relacionado afecta a una propuesta reciente de Eric Drexler (en <http://www.fhi.ox.ac.uk/reports/2015-3.pdf>) para mantener la superinteligencia bajo control compartimentándola en partes simples, ninguna de las cuales comprendería la situación en su conjunto. Si esto funciona, tendríamos de nuevo una herramienta extraordinariamente potente sin escrúpulos morales intrínsecos, que implementaría todos los caprichos de su propietario sin ningún reparo. Recordaría a la burocracia compartimentada de una dictadura distópica: una parte construye armas sin saber cómo serán utilizadas, otra ejecuta a los presos sin saber por qué fueron condenados, y así

sucesivamente.
[97] Una variante moderna de la regla de oro es la idea, debida a John Rawls, de que una situación hipotética es justa si nadie la cambiaría sin saber de antemano cuál de las personas en dicha situación le tocaría ser.
[98] Por ejemplo, se ha sabido que el cociente intelectual de muchos de los altos mandos de Hitler era bastante alto. Véase «How Accurate Were the IQ Scores of the High-Ranking Third Reich Officials Tried at Nuremberg?», *Quo*ra, disponible online en <http://tinyurl.com/nurembergiq>.

CONSCIENCIA

[99] La entrada sobre la consciencia, escrita por Stuart Sutherland, es bastante divertida*: Macmillan Dictionary of Psycholo*gy (Londres, Macmillan, 1989).
[100] Erwin Schrödinger, uno de los padres fundadores de la mecánica cuántica, hizo este comentario en su libr*o Mente y materia*, al reflexionar sobre el pasado y sobre lo que podría haber sido si nunca hubiese surgido la vida consciente. Por otra parte, la irrupción de la IA plantea la posibilidad lógica de que en el futuro se acabe representando una obra ante un auditorio de asientos vacíos.
[101] L*a Stanford Encyclopedia of Philosop*hy incluye una extensa lista de las distintas definiciones y usos de la palabra «consciencia», en <http://tinyurl.com/stanfordconsciousness>.
[102] Yuval Noah Harari, *Homo Deus: A Brief History of Tomorr*ow (Nueva York, Harper-Collins, 2017), p. 116 [hay trad. cast.*: Homo Deus. Una breve historia del maña*na, Barcelona, Debate, 2016].
[103] Una excelente introducción a los sistemas 1 y 2, escrita por uno de los pioneros en su estudio es: Daniel Kahneman, *Thinking, Fast and Sl*ow (Nueva York, Farrar, Straus & Giroux, 2011) [hay trad. cast.*: Pensar rápido, pensar despac*io, Barcelona, Debate, 2015].
[104] Véase Christof Koch*, The Quest for Consciousness: A Neurobiological Approa*ch (Nueva York, W. H. Freeman, 2004).
[105] Puede que solo seamos conscientes de una minúscula proporción (entre 10 y 50 bits) de la información que llega a nuestro cerebro cada segundo: K. Küpfmüller, 1962, «Nachrichtenverarbeitung im Menschen», e*n Taschenbuch der Nachrichtenverarbeitu*ng, K. Steinbuch, ed. (Berlín, Springer-Verlag, 1962), pp. 1481-1502. T. Nørretranders, *The User Illusion: Cutting Consciousness Down to Si*ze (Nueva York, Viking, 1991).
[106] Michio Kaku, *The Future of the Mind: The Scientific Quest to Understand, Enhance, and Empower the Mi*nd (Nueva York, Doubleday, 2014) [hay trad. cast.*: El futuro de nuestra mente. El reto científico para entender, mejorar, y fortalecer nuestra men*te, Barcelona, Debolsillo, 2017]; Jeff Hawkins y Sandra Blakeslee*, On Intelligen*ce (Nueva York, Times Books, 2007) [hay trad. cast.*: Sobre la inteligen*cia, Barcelona, Espasa, 2005]; Stanislas Dehaene, Michel Kerszberg y Jean-Pierre Changeux, «A Neuronal Model of a Global Workspace in Effortful Cognitive Tasks»*, Proceedings of the National Academy of Scienc*es 95 (1998), pp. 14529-14534.
[107] Vídeo que celebra el famoso experimento de «huele a tostada quemada» de Penfield, en <https://www.youtube.com/watch?v=mSN86kphL68>. Detalles sobre la corteza sensomotora: Elaine Marieb y Katja Hoehn*, Anatomy & Physiolo*gy, 3.ª ed. (Upper Saddle River, Nueva Jersey, Pearson, 2008), pp. 391-395.
[108] En los últimos años, el estudio de los correlatos neuronales de la conciencia (CNC) ha pasado a ser algo bastante común en la comunidad neurocientífica. Véanse, por ejemplo, Geraint Rees, Gabriel Kreiman y Christof Koch, «Neural Correlates of Consciousness in Humans»*, Nature Reviews Neuroscien*ce 3 (2002), pp. 261-270, y Thomas Metzinger, *Neural Correlates of Consciousness: Empirical and Conceptual Questio*ns (Cambridge, Massachusetts, MIT Press, 2000).

[109] Cómo funciona la supresión continua con flash: Christof Koch, *The Quest for Consciousness: A Neurobiological Approa*ch (Nueva York, W. H. Freeman, 2004); Christof Koch y Naotsugu Tsuchiya, «Continuous Flash Suppression Reduces Negative Afterimages», *Nature Neuroscien*ce 8 (2005), pp. 1096-1101.
[110] Christof Koch, Marcello Massimini, Melanie Boly y Giulio Tononi, «Neural Correlates of Consciousness: Progress and Problems», *Nature Reviews Neuroscien*ce 17 (2016), p. 307.
[111] Véase Koch, *The Quest for Consciousne*ss, p. 260. Para más información, también puede consultarse l*a Stanford Encyclopedia of Philosop*hy, en <http://tinyurl.com/consciousnessdelay>.
[112] Sobre la sincronización de la percepción consciente: David Eagleman, *The Brain: The Story of Y*ou (Nueva York, Pantheon, 2015) [hay trad. cast.*: El cerebro. Nuestra histor*ia, Barcelona, Anagrama, 2017] *y Stanford Encyclopedia of Philosop*hy, en <http://tinyurl.com/consciousnesssync>.
[113] Benjamin Libet, *Mind Time: The Temporal Factor in Consciousne*ss (Cambridge, Massachusetts, Harvard University Press, 2004); Chun Siong Soon, Marcel Brass, Hans-Jochen Heinze y John-Dylan Haynes, «Unconscious Determinants of Free Decisions in the Human Brain», *Nature Neuroscien*ce 11 (2008), pp. 543-545, disponible online en <http://www.nature.com/neuro/journal/v11/n5/full/nn.2112.html>.
[114] Ejemplos de enfoques teóricos recientes sobre la consciencia:

- Daniel Dennett, *Consciousness Explain*ed (Back Bay Books, 1992) [hay trad. cast.*: La conciencia explicada. Una teoría interdisciplin*ar, Barcelona, Paidós Ibérica, 1995].
- Bernard Baars, *In the Theater of Consciousness: The Workspace of the M*ind (Nueva York, Oxford University Press, 2001).
- Christof Koch, *The Quest for Consciousness: A Neurobiological Approa*ch (Nueva York, W. H. Freeman, 2004).
- Gerald Edelman y Giulio Tononi, *A Universe of Consciousness: How Matter Becomes Imaginati*on (Nueva York, Hachette, 2008) [hay trad. cast.*: El universo de la conciencia. Cómo la materia se convierte en imaginaci*ón, Barcelona, Crítica, 2002.
- António Damásio, *Self Comes to Mind: Constructing the Conscious Bra*in (Nueva York, Vintage, 2012) [hay trad. cast.*: Y el cerebro creó al hombre: ¿Cómo pudo el cerebro generar emociones, sentimientos, ideas y el y*o?, Barcelona, Crítica, 2010].
- Stanislas Dehaene, *Consciousness and the Brain: Deciphering How the Brain Codes Our Though*ts (Nueva York, Viking, 2014) [hay trad. cast.*: La conciencia en el cerebro. Descifrando el enigma de cómo el cerebro elabora nuestros pensamient*os, Buenos Aires, Siglo Veintiuno Editores Argentina, 2015].
- Stanislas Dehaene, Michel Kerszberg and Jean-Pierre Changeux, «A Neuronal Model of a Global Workspace in Effortful Cognitive Tasks», *Proceedings of the National Academy of Scienc*es 95 (1998), pp. 14529-14534.
- Stanislas Dehaene, Lucie Charles, Jean-Rémi King y Sébastien Marti, «Toward a Computational Theory of Conscious Processing», *Current Opinion in Neurobiolo*gy 25 (2014), pp. 760-784.

[115] Detallada discusión de distintos usos del término «emergencia» en física y en filosofía por David Chalmers, en <http://cse3521.artifice.cc/Chalmers-Emergence.pdf>.
[116] Aquí argumento que la consciencia es la manera en que se siente la información cuando se procesa de determinadas maneras complejas, en <https://arxiv.org/abs/physics/0510188>, <https://arxiv.org/abs/0704.0646>, Max Tegmark, *Our Mathematical Univer*se (Nueva York, Knopf, 2014) [hay trad. cast.*: Nuestro universo matemát*ico, Barcelona, Antoni Bosch, 2015]. David Chalmers expresa una opinión similar en su libro de 199*6 The Conscious Mi*nd [hay trad. cast.*: La mente conscien*te, Barcelona, Gedisa, 2013]: «La experiencia es la información desde el interior; la física es la información desde el exterior».
[117] Adenauer Casal*i et a*l., «A Theoretically Based Index of Consciousness Independent of Sensory

Processing and Behavior», *Science Translational Medicine* 5 (2013), p. 198ra105, en <http://tinyurl.com/zapzip>.

[118] La teoría de la información integrada no es válida en sistemas continuos.

- <https://arxiv.org/abs/1401.1219>
- <http://journal.frontiersin.org/article/10.3389/fpsyg.2014.00063/full>
- <https://arxiv.org/abs/1601.02626>

[119] Entrevista con Clive Wearing, cuya memoria a corto plazo es de tan solo unos 30 segundos, en <https://www.youtube.com/watch?v=WmzU47i2xgw>.

[120] La crítica de la TII de Scott Aaronson, en < http://www.scottaaronson.com/blog/?p=1799>.

[121] La crítica de la TII de Cerrullo, que argumenta que la integración no es una condición suficiente para la consciencia, en <http://tinyurl.com/cerrullocritique>.

[122] La predicción de la TII de que los humanos simulados serán zombis, en <http://rstb.royalsocietypublishing.org/content/370/1668/20140167>.

[123] La crítica de la TII de Shanahan, en < https://arxiv.org/pdf/1504.05696.pdf>.

[124] Blindsight: <http://tinyurl.com/blindsight-paper>.

[125] Puede que solo seamos conscientes de una minúscula proporción (digamos que entre 10 y 50 bits) de la información que nuestro cerebro recibe cada segundo: Küpfmüller, «Nachrichtenverarbeitung im Menschen»; Nørretranders, *The User Illusion*.

[126] Argumento a favor y en contra de la «consciencia sin acceso»: Victor Lamme, «How Neuroscience Will Change Our View on Consciousness», *Cognitive Neuroscience* (2010), pp. 204-220, disponible online en <http://www.tandfonline.com/doi/abs/10.1080/17588921003731586>.

[127] «Selective Attention Test», en <https://www.youtube.com/watch?v=vJG698U2Mvo>.

[128] Véase Lamme, «How Neuroscience Will Change Our View on Consciousness», n.º 28.

[129] Esta y otras cuestiones relacionadas se discuten en detalle en el libro de Daniel Dennett *La conciencia explicada*.

[130] Véase Kahneman, *Pensar rápido, pensar despacio*, citado en nota n.º 5.

[131] La *Stanford Encyclopedia of Philosophy* revisa la controversia en torno al libre albedrío, en <https://plato.stanford.edu/entries/freewill>.

[132] Vídeo en el que Seth Lloyd explica por qué una IA sentirá que tiene libre albedrío, en <https://www.youtube.com/watch?v=Epj3DF8jDWk>.

[133] Véase Steven Weinberg, *Dreams of a Final Theory: The Search for the Fundamental Laws of Nature* (Nueva York, Pantheon, 1992) [hay trad. cast.: *El sueño de una teoría final. La búsqueda de las leyes fundamentales de la naturaleza*, Barcelona, Booket, 2010].

[134] El primer análisis científico concienzudo de nuestro futuro remoto: Freeman J. Dyson, «Time Without End: Physics and Biology in an Open Universe», *Reviews of Modern Physics* 51, n.º 3 (1979), p. 447, disponible online en <http://blog.regehr.org/extra_files/dyson.pdf>.

EPÍLOGO

[135] La carta abierta (en <http://futureoflife.org/ai-open-letter>) que resultó de la conferencia de Puerto Rico sostenía que la investigación sobre cómo hacer que los sistemas de IA sean potentes y beneficiosos es importante y oportuna, y que hay líneas de investigación concretas que pueden explorarse a día de hoy, como se pone de manifiesto en este documento sobre prioridades en la investigación, en <http://futureoflife.org/data/documents/research_priorities.pdf>.

[136] Vídeo de mi entrevista con Elon Musk sobre seguridad en IA que puede verse en YouTube, en <https://www.youtube.com/watch?v=rBw0eoZTY-g>.

[137] Una buena compilación de vídeos de casi todos los intentos de aterrizaje de cohetes de SpaceX,

que culminan con el primer aterrizaje con éxito en el mar, en <https://www.youtube.com/watch?v=AllaFzIPaG4>.

[138] Tuit de Elon Musk sobre nuestro concurso de becas en IA segura, en <https://twitter.com/elonmusk/status/555743387056226304>.

[139] Tuit de Elon Musk sobre nuestra carta abierta en apoyo de la IA segura, en <https://twitter.com/elonmusk/status/554320532133650432>.

[140] En «An Open Letter to Everyone Tricked into Fearing Artificial Intelligence» *(Popular Scien*ce, 14 de enero de 2015), Erik Sofge se burlaba de la cobertura alarmista que nuestra carta abierta recibió en los medios, en <http://www.popsci.com/open-letter-everyone-tricked-fearing-ai>.

[141] Tuit de Elon Musk sobre su cuantiosa donación al Future of Life Institute y sobre el mundo de los investigadores en IA segura, en <https://twitter.com/elonmusk/status/555743387056226304>.

[142] Para más información sobre el consorcio Partnership on Artificial Intelligence to Benefit People and Society, véase su sitio web, en <https://www.partnershiponai.org>.

[143] Ejemplos de informes recientes que expresan opiniones sobre IA: One Hundred Year Study on Artificial Intelligence, informe del grupo de estudio de 2015, «Artificial Intelligence and Life in 2030» (septiembre de 2016), en <http://tinyurl.com/stanfordai>; informe de la Casa Blanca sobre el futuro de la IA, en <http://tinyurl.com/obamaAIreport>; informe de la Casa Blanca sobre IA y empleo, en <http://tinyurl.com/AIjobsreport>; informe del IEEE sobre IA y bienestar humano, «Ethically Aligned Design, Version 1» (13 de diciembre de 2016), en <http://standards.ieee.org/develop/indconn/ec/ead_v1.pdf>; hoja de ruta para U.S. Robotics, en <http://tinyurl.com/roboticsmap>.

[144] Entre los principios que no superaron el corte final, uno de mis favoritos era este: «Precaución con la consciencia: puesto que no hay consenso al respecto, deberíamos evitar hacer suposiciones fuertes sobre si la inteligencia artificial avanzada poseerá o requerirá consciencia o sentimientos». Pasó por muchas iteraciones y, en la última, la controvertida palabra «consciencia» se sustituyó por «experiencia subjetiva», pero este principio obtuvo solo el 88 % de aprobación, apenas por debajo del umbral del 90 %.

[145] Mesa redonda sobre superinteligencia con Musk y otros grandes cerebros, en <http://tinyurl.com/asilomarAI>.

NOTAS EXPLICATIVAS

(1) Por cuestiones de simplicidad, para esta historia he supuesto que la economía y la tecnología serían las actuales, aunque la mayoría de los investigadores aventuran que aún quedan al menos varias décadas para que se alcance una IA general de nivel humano. El plan de los omegas debería ser aún más fácil de llevar a cabo con éxito en el futuro si la economía digital continúa creciendo y son cada vez más los servicios que pueden solicitarse a través de internet sin traba alguna.

(2) ¿Por qué la vida se fue haciendo más compleja? La evolución premia a la vida lo suficientemente compleja para predecir y sacar provecho de las regularidades existentes en su entorno, por lo que en un entorno más complejo evolucionará una vida más compleja e inteligente. Esta vida más inteligente crea un entorno más complejo para las otras formas de vida con las que compite, y estas a su vez evolucionan para ser más complejas, lo cual acaba dando lugar a un ecosistema de vida muy complejo.

(3) La conversación en torno a la IA es importante tanto por lo que respecta a su urgencia como a su impacto. En cuanto al cambio climático, que podría causar estragos en cincuenta o doscientos años, muchos expertos vaticinan que la IA tendrá un impacto mucho mayor dentro de algunas décadas, y que podría proporcionarnos la tecnología con la que mitigar ese cambio. Respecto de las cuestiones relacionadas con las guerras, el terrorismo, el desempleo, la pobreza, las migraciones y la justicia social, la irrupción de la IA tendrá un impacto general mayor. De hecho, a lo largo del libro veremos cómo la IA puede llegar a determinar lo que sucederá con todos estos asuntos.

(4) Para entenderlo, imaginemos cómo reaccionaríamos si alguien afirmase que la capacidad para alcanzar logros de nivel olímpico puede cuantificarse a través de un solo número llamado «cociente atlético» (CA), tal que el deportista olímpico con el mayor CA ganaría la medalla de oro en todos los deportes.

(5) Hay quien prefiere usar «IA de nivel humano» o «IA fuerte» como sinónimos de IAG, pero ambas presentan problemas. Incluso una calculadora de bolsillo es una IA de nivel humano en el sentido estrecho. El antónimo de «IA fuerte» es «IA débil», pero se hace raro decir que sistemas de IA estrecha como Deep Blue, Watson y AlphaGo son «débiles».

(6) NAND es una contracción de NOT AND: una puerta AND devuelve 1 solo si la primera entrada es 1 y la segunda es 1, de manera que una NAND devuelve exactamente lo opuesto.

(7) Aquí, «función bien definida» equivale a lo que los matemáticos e informáticos llaman «función computable», esto es, una función que podría ser computada por un hipotético ordenador con memoria y tiempo ilimitados. Como es bien sabido, Alan Turing y Alonzo Church demostraron que también existen funciones que pueden describirse pero que no son computables.

(8) Si le gustan las matemáticas, sepa que dos de las opciones más populares para esta función de activación son la llamada función sigmoide, $\sigma(x) = 1/(1 + e^{-x})$, y la función rampa, $\sigma(x) = \max\{0, x\}$, aunque se ha demostrado que casi cualquier función valdría siempre que no sea lineal (una línea recta). El famoso modelo de Hopfiel usa $\sigma(x) = -1$ si $x < 0$ y $\sigma(x) = 1$ si $x \geq 0$. Si los estados de las neuronas se almacenan en un vector, entonces la red se actualiza simplemente multiplicando ese vector por una matriz que contiene los acoplamientos sinápticos y aplicando a continuación la función σ a todos los elementos.

(9) Si quiere ver un mapa más detallado del panorama de la investigación en IA segura, aquí hay uno

interactivo, desarrollado gracias a un esfuerzo comunitario liderado por Richard Mallah del FLI: <https://futureoflife.org/landscape>.

(10) Dicho con más precisión: la verificación pregunta si un sistema satisface sus especificaciones, mientras que la validación pregunta si se eligieron las especificaciones correctas.

(11) Aun incluyendo esta colisión en las estadísticas, se ha visto que, cuando se activa, el piloto automático de Tesla reduce los accidentes en un 40 %: <http://tinyurl.com/teslasafety>.

(12) Recordemos que las FLOPS son las operaciones de coma flotante por segundo; por ejemplo, cuántos números de 19 dígitos pueden multiplicarse cada segundo.

(13) Como ha explicado Bostrom, la capacidad de simular a un destacado investigador humano en IA a un coste mucho menor que su sueldo por hora permitiría a una compañía de IA incrementar drásticamente la capacidad de su fuerza laboral, acumulando una enorme riqueza y acelerando recursivamente sus avances en la construcción de mejores ordenadores y, en última instancia, mentes más inteligentes.

(14) La idea se remonta a san Agustín, que dejó escrito que «si algo no mengua al compartirlo con otros, no se posee correctamente si solo se posee y no se comparte».

(15) El primero en sugerirme esta idea fue mi amigo y colega Anthony Aguirre.

(16) El prestigioso cosmólogo Fred Hoyle exploró un escenario relacionado con un enfoque diferente en la serie televisiva británica *A de Andrómeda*.

(17) Inyectar carbono en la atmósfera puede provocar dos tipos de cambio climático: calentamiento debido al dióxido de carbono o enfriamiento por el humo y el hollín. No es solo el primer tipo el que ocasionalmente se descarta sin evidencia científica: he oído decir que el invierno nuclear ha sido desacreditado y es prácticamente imposible. Siempre respondo pidiendo una referencia a un artículo científico revisado por pares que haga afirmaciones tan contundentes, y hasta ahora parece que no hay ninguno. Aunque existe una gran incertidumbre que justifica una mayor investigación, especialmente en relación con la cantidad de humo que se produciría y hasta dónde subiría, no hay en mi opinión científica una base real para descartar el riesgo de invierno nuclear.

(18) Si usted trabaja en el sector energético, quizá esté acostumbrado a definir la eficiencia como la proporción útil del total de energía liberada.

(19) Si no se encuentra un agujero negro de origen natural en el universo cercano, se puede crear uno nuevo metiendo una cantidad enorme de materia en un espacio lo suficientemente pequeño.

(20) Esta es una simplificación algo excesiva, porque la radiación de Hawking también incluye algunas partículas de las cuales es difícil extraer trabajo útil. Los agujeros negros grandes tienen una eficiencia de solo 90 %, porque aproximadamente el 10 % de su energía se irradia en forma de gravitones: partículas muy tímidas que son casi imposibles de detectar, y más difícil aún es extraer de ellas trabajo útil. A medida que el agujero negro va evaporándose y decreciendo, la eficiencia disminuye aún más porque la radiación de Hawking comienza a incluir neutrinos y otras partículas masivas.

(21) Para los lectores que sean seguidores de Douglas Adams, senalaré que esta es una cuestión elegante que da respuesta a la pregunta sobre la vida, el universo y todo lo demás. Más exactamente, la eficiencia es $1 - 1\sqrt{3} \approx 42\ \%$.

(22) Si alimentamos el agujero negro colocando a su alrededor una nube de gas que gire lentamente en la misma dirección, entonces este gas girará cada vez más rápido a medida que sea atraído y engullido por el agujero, lo cual acelerará la rotación de este, del mismo modo que un patinador gira más rápido cuando recoge los brazos. Esto puede hacer que el agujero siga girando a la velocidad máxima, lo que permite extraer el 42 % de la energía del gas y luego el 29 % del resto, hasta alcanzar una eficiencia total del $42\ \% + (1 - 42\ \%) \times 29\ \% \approx 59\ \%$.

(23) Debe alcanzar una temperatura suficientemente elevada para reunificar las fuerzas electromagnética y débil, lo cual sucede cuando las partículas se mueven a la velocidad a la que son aceleradas por 200 mil millones de voltios en un colisionador de partículas.

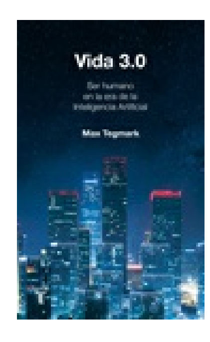
(24) Más arriba solo hemos hablado de la materia hecha de átomos. Hay alrededor de seis veces más materia oscura, pero es muy difícil de detectar y normalmente atraviesa la Tierra de un lado a otro sin verse afectada por nada, por lo que aún nos queda averiguar si es posible capturarla y utilizarla en el futuro.
(25) Los cálculos cósmicos resultan ser sorprendentemente simples: si la civilización se expande a través del espacio en expansión no a la velocidad de la luz *c*, sino a otra velocidad inferior *v*, el número de galaxias colonizadas se reduce en un factor $(v/c)^3$. Esto significa que las civilizaciones rezagadas son severamente penalizadas: una civilización que se expanda diez veces más lentamente acabaría colonizando en última instancia un número de galaxias mil veces menor.
(26) De ahí proviene el nombre de «Big Snap», literalmente «Gran Rotura». *(N. del T.)*
(27) Sin embargo, John Gribbin llega a una conclusión similar en su libro de 2011 *Solos en el universo*. Para una variedad de perspectivas sugerentes sobre esta cuestión, recomiendo también el libro de Paul Davies *Un silencio inquietante*, publicado también en 2011.
(28) Para decidir hacia dónde dirigirse, muchos insectos usan una regla empírica consistente en suponer que un foco de luz intensa es el Sol y en volar formando un determinado ángulo fijo respecto a él. Por desgracia, si resulta que la luz es una llama cercana, este truco puede fallar y llevar al insecto a una mortífera espiral hacia la llama.
(29) Utilizo la expresión «mejorar su software» en el sentido más amplio, que incluye no solo optimizar sus algoritmos sino también hacer que su proceso de toma de decisiones sea más racional, de manera que alcance la mayor capacidad posible de lograr sus objetivos.
(30) Un punto de vista alternativo es el del *dualismo cartesiano*, según el cual las entidades vivas se diferencian de las inanimadas en que contienen una sustancia no física como el «ánima», el «élan vital» o el «alma». El respaldo hacia el dualismo cartesiano entre los científicos ha ido disimuyendo progresivamente. Para entender por qué, tengamos en cuenta que nuestro cuerpo está compuesto por alrededor de 10^{29} quarks y electrones, que, hasta donde sabemos, se mueven de acuerdo con unas leyes físicas simples. Imaginemos una tecnología futura capaz de seguir la pista de todas nuestras partículas: si se comprobase que obedecen exactamente las leyes físicas, nuestra supuesta alma no estaría teniendo ningún efecto sobre nuestras partículas, por lo que nuestra mente consciente y nuestra capacidad de controlar nuestros movimientos no tendrían nada que ver con un alma. Si, por el contrario, se descubriese que las partículas no obedecen las leyes de la física conocidas porque nuestra alma las empujaba de un sitio a otro, la nueva entidad causante de estas fuerzas sería, por definición, una entidad física que podríamos estudiar como hicimos en el pasado con nuevos campos y partículas.
(31) Utilizo la palabra «qualia» de acuerdo con la definición de Wikipedia, para referirme a instancias individuales de experiencia subjetiva, es decir, a la experiencia subjetiva en sí misma, no a ninguna sustancia supuesta que la cause. Tenga en cuenta que algunas personas usan la palabra de manera diferente.
(32) Si nuestra realidad física es enteramente matemática (basada en información, a grandes rasgos), como analicé en mi libro *Nuestro universo matemático*, entonces ningún aspecto de la realidad —ni siquiera la consciencia— está fuera del ámbito de la ciencia. De hecho, el problema realmente difícil de la consciencia es, desde ese punto de vista, exactamente el mismo problema que el de comprender cómo algo matemático puede sentirse como físico: si parte de una estructura matemática es consciente, experimentará el resto de la misma como su mundo físico exterior.
(33) Aunque en anteriores ocasiones usé «perceptronio» como sinónimo de sentronio, ese nombre sugiere una definición demasiado limitada, ya que las percepciones son simplemente aquellas experiencias subjetivas que percibimos a partir de la información sensorial, lo que excluye, por ejemplo, los sueños y los pensamientos generados de forma interna.
(34) Existe una posible tensión entre esta afirmación y la idea de que la consciencia es independiente

del sustrato, puesto que, aunque el procesamiento de información puede ser diferente al nivel más bajo, es por definición idéntico a los niveles más altos, donde determina el comportamiento.

(35) Esta experiencia también me hizo replantearme cómo interpretar las noticias. Aunque, obviamente, ya antes era consciente de que la mayoría de los medios tienen su propia línea política, me di cuenta entonces de que también tienen sesgos que los alejan del centro en todas las cuestiones, incluso las que no son políticas.

¿Cómo afectará la inteligencia artificial al crimen, a la guerra, a la justicia, al trabajo, a la sociedad y al sentido de nuestras vidas?

Bienvenidos a la conversación más importante de nuestro tiempo.



¿Cómo afectará la inteligencia artificial al crimen, a la guerra, a la justicia, al trabajo, a la sociedad y al sentido de nuestras vidas? ¿Es posible que las máquinas nos dejen fuera de juego, remplazando a los humanos en el mercado laboral e incluso en otros ámbitos? ¿La inteligencia artificial proveerá mejoras sin precedente a nuestras vidas o nos dará más poder del que podemos manejar?

Muchas de las cuestiones más fundamentales de la actualidad están íntimamente relacionadas con el aumento de la inteligencia artificial.

Max Tegmark no se asusta ante la gama completa de puntos de vista o ante los temas más controvertidos, desde la superinteligencia hasta el significado, la conciencia y los límites físicos últimos de la vida en el cosmos. En ***Vida 3.0***, clarifica los conceptos clave necesarios para hablar de inteligencia artificial al tiempo que ayuda a entender la importancia de las cuestiones clave, aquellas que la humanidad tendrá que abordar en las próximas décadas.

Reseñas:
«Todos nosotros, no solo científicos, industriales y generales, deberíamos preguntarnos qué puede hacerse ahora para aumentar las posibilidades de cosechar los beneficios de la IA futura y evitar sus riesgos. Esta es la conversación más importante de nuestro tiempo, y con este estimulante libro Tegmark te ayudará a participar en ella.»
Stephen Hawking

«Enriquecedor y visionario. Todo el mundo debería leerlo.»

The Times

«Tegmark explica con brillantez numerosos conceptos del terreno de la informática al de la cosmología, escribe con modestia y sutileza intelectual, le ofrece al lector el importante servicio de definir sus términos con claridad, y con razón rinde homenaje a las mentes creativas de los escritores de ciencia ficción que, por supuesto, abordaron este tipo de preguntas hace más de medio siglo.»
Steven Poole, *The Telegraph*

«Original, accesible y provocador. Tegmark ilumina las numerosas facetas de la inteligencia artificial. Disfruten del viaje y saldrán del otro extremo con una mejor apreciación de adónde podría llevarnos la tecnología en los próximos años.»
Science

«Una guía convincente por los retos y dilemas en nuestra búsqueda de un gran futuro de la vida, la inteligencia y la consciencia, en la Tierra y más allá de esta.»
Elon Musk

Estimulante. La discusión inteligente y desenfrenada de Tegmark conduce a fascinantes especulaciones sobre civilizaciones basadas en la inteligencia artificial que abarcan galaxias y eones. Absorbente.»
Publishers Weekly

SOBRE EL AUTOR

Max Tegmark (Suecia, 1967) es profesor de Física en el MIT, director científico del Foundational Questions Institute y cofundador del Future of Life Institute. Ha publicad*o Nuestro universo matemáti*co y, en 2007, *Forb*es lo eligió como una de las «diez personas que podrían cambiar el mundo».

Título original: *Life 3.0*

Travessera de Gràcia, 47-49. 08021 Barcelona

ISBN ebook: 978-84-306-1965-8
Diseño de la cubierta: Penguin Random House Grupo Editorial adaptación del diseño original de Penguin UK
Imagen de la cubierta: Getty Images
Conversión ebook: Arca Edinet S. L.

www.megustaleer.com



ÍNDICE